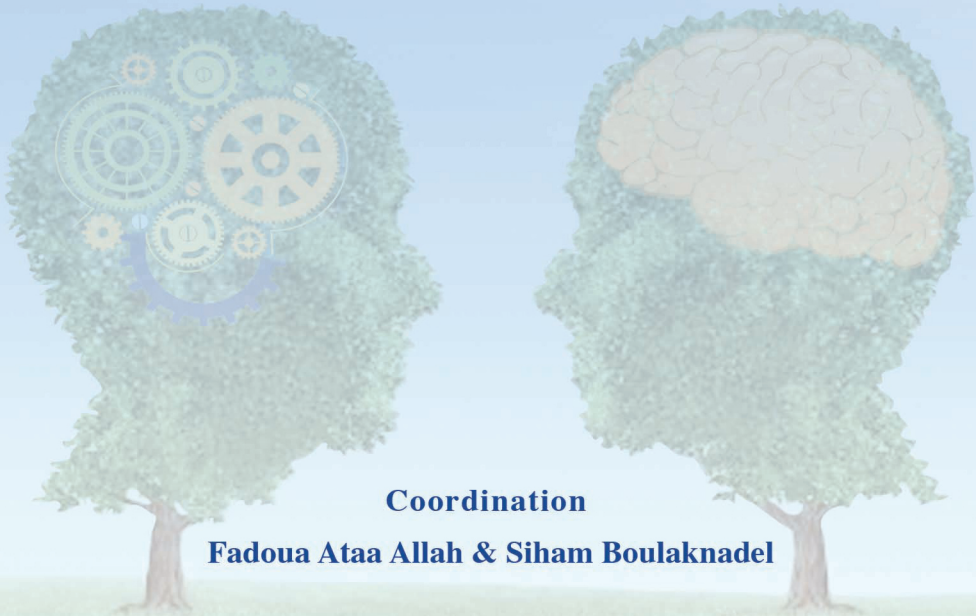
Diversité & TAL Linguistique

Actes de la 2e conférence internationale



Coordination

Fadoua Ataa Allah & Siham Boulaknadel

Centre des Etudes Informatiques, des Systèmes
d'Information et de Communication

DiLiTAL'2019

Diversité Linguistique & TAL

16 et 17 décembre 2019

ACTES DE LA 2e EDITION DE LA CONFERENCE INTERNATIONALE

Coordination

Fadoua ATAA ALLAH & Siham BOULAKNADEL

Publications de l'Institut Royal de la Culture Amazighe
Centre des Etudes Informatiques, des Systèmes d'Information et de Communication

Série : **Colloques et séminaires** ***N° 68***

Titre
DIVERSITE LINGUISTIQUE ET TAL
Actes de la 2e édition de la conférence internationale

Coordination
Fadoua ATAA ALLAH & Siham BOULAKNADEL

Editeur
Institut Royal de la Culture Amazighe

Imprimerie
Edition & Impression Bouregreg - Rabat

Dépôt légal

ISBN

ISSN

Copyright
©IRCAM

PREFACE

A la fin de l'atelier DiLiTAL, qui s'est tenu le 26 juin 2017 à Orléans, lors de la conférence JEP-TALN RECITAL 2017, il a été proposé une continuité de cet atelier sous forme de conférence. Ainsi, la 2e édition s'est tenue les 16 et 17 décembre 2019 à Oujda, en collaboration avec l'Université Mohamed Premier Oujda.

Les recherches en traitement automatique des langues 'mineures/minorées' sont actuellement à l'orée de développements majeurs. Les efforts de reconnaissance de ces langues et de standardisation des différents alphabets commencent à porter leurs fruits. Pour autant, la plupart de ces langues sont peu dotées informatiquement, notamment les ressources électroniques disponibles sont rares, mal distribuées, voire inexistantes. Seules sont accessibles les fonctions d'édition et d'impression rendant l'exploitation de ces langues difficiles. De ce fait, le développement de ces langues s'impose comme une nécessité vitale du moment où il est question de développer leur usage à l'écrit et à l'oral dans l'administration, la vie quotidienne et surtout de créer des normes d'écriture standardisées dans un souci de les introduire dans le système éducatif.

Conscients de cette situation, la conférence DiLiTAL'2019 s'est donnée pour objectif de (1) favoriser le traitement automatique des langues dans un contexte plurilingue et de sensibiliser la communauté scientifique aux enjeux et aux difficultés rencontrées dans le cas du traitement des Langues Peu Dotées (LPD) ; (2) montrer l'intérêt de créer des outils génériques permettant de traiter un très grand nombre de langues, tout en identifiant les besoins spécifiques à chaque langue ; (3) donner aux chercheurs qui s'intéressent au Traitement Automatique des Langues (TAL) l'opportunité de présenter les différentes méthodes et techniques utilisées pour dynamiser la construction et la mutualisation des ressources, ainsi que le transfert des savoirs et des savoir-faire.

Ainsi, la conférence DiLiTAL'2019 a ouvert ses portes aux contributions présentant des interventions théoriques ou des applications concrètes qui s'articulent autour des points suivants :

1. Ressources et corpus : production, standardisation et archivage

L'ambition de cet axe est d'explorer les initiatives actuelles et futures qui ont pour but de collecter et de structurer des données langagières (spécialisées ou généralistes) des LPD. Un intérêt particulier est accordé aux questions d'accessibilité et de portabilité des données linguistiques qui s'avèrent être des problèmes majeurs dans les travaux consacrés aux LPD. Par conséquent,

les contributions concernant la création des ressources open source ont été particulièrement encouragées.

2. Outils pour le traitement des LPD

Les interrogations de ce second axe portent sur la pertinence de l'utilisation d'outils existants pour les LPD et sur la manière dont ils gèrent le multilinguisme. En effet, depuis l'arrivée d'UTF-8 (Unicode), le TAL s'est doté de la possibilité de diversifier son terrain d'action-recherche, permettant de traiter d'autres langues que celles dites « majoritaires » (et bien dotées).

3. Questions sociales, culturelles et éthiques

DiLiTAL a été également une occasion d'examiner les possibilités de coopérations transfrontalières pour encourager l'échange d'expériences et dynamiser la recherche en TAL dans les pays où les LPD sont en usage. Par ailleurs, une réflexion a été entamée concernant les aspects éthiques de la collecte, de l'archivage et du traitement des données linguistiques.

4. Retours d'expérience

Certaines contributions ont mis en lumière des expériences de traitement des LPD. La synthèse d'expertises acquises par des chercheurs ayant participé à ce genre de travaux a permis de nourrir la réflexion épistémologique nécessaire au développement des outils pour dépasser le cadre d'un seul type de langue.

Nos remerciements chaleureux vont tout d'abord aux auteurs pour leurs contributions scientifiques. Nous remercions aussi les membres du comité de lecture, pour le temps consacré, pour leur patience et leur bonne volonté.
Nos vifs remerciements vont également à toute l'équipe du comité d'organisation qui a mis tout en œuvre pour pouvoir permettre le meilleur accueil et le meilleur déroulement de la conférence. Enfin, nous remercions spécialement pour son soutien financier et aides diverses l'Institut Royal de la Culture Amazighe, l'Université Mohamed Premier d'Oujda et le Laboratoire de Recherche en Informatique. Sans ces soutiens, ni la conférence DiLiTAL 2019, ni ce recueil n'auraient vu le jour.

Comité scientifique

Hamouda Bebaha (Institut Doha, Qatar)
Delphine Bernhard (LILPA, France)
Laurent Besacier (LIG, France)
Violetta Cavalli-Sforza (UAI, Maroc)
Khalid Choukri (ELDA, France)
Karim Darwish (Univ. Qatar, Qatar)
Hammou Fadili (CNAM, France)
Ahmed Guessoum (USTHB, Algérie)
Kais Haddar (FSS, Tunisie)
Najlae Idrissi (FST-Béni Mellal, Maroc)
Ali Mohamed Jaoua (Univ. Qatar, Qatar)
Mustafa Jarrar (Univ. Birzeit, Palestine)
Joseph Mariani (LIMSI, France)
Denis Maurel (LI, France)
Rochdi Messoussi (FS-Kénitra, Maroc)
Emmanuel Morin (Univ-Nantes, France)
Patrice Pognan (PLIDAM, France)
Khalid Satori (USMBA, Maroc)
Hassan Satori (FP-Nador, Morocco)
Majdi Sawalha (Univ. Jordanie, Jordanie)
Khaled Shaalan (BUD, Dubai)
Yahya Slimani (Univ. Manouba, Tunisie)
Nora Tigziri (UMMTO, Algérie)
Farida Yamouni (UMMTO, Algérie)
Imad Zeroual (Univ. Moulay Ismail, Maroc)
Michael Zock (LIF, France)
Mounir Zrigui (Univ. Monastir, Tunisie)

Coordinateurs et Comité d'organisation

Coordinateurs

Fadoua Ataa Allah (IRCAM, Maroc)
Azzedine Mazroui (FS-Oujda, Maroc)

Comité d'organisation

Abdelhak Lakhouaja (FS-Oujda, Maroc)
Siham Boulaknadel (IRCAM, Maroc)
Abdelouafi Meziane (FS-Oujda, Maroc)
Youness Chaabi (IRCAM, Maroc)
Safae Berrichi (FS-Oujda, Maroc)
Naoual Nassiri (FS-Oujda, Maroc)
Ibtissam Touahri (FS-Oujda, Maroc)
Omaima Zine (FS-Oujda, Maroc)
Jamal Frain (IRCAM, Maroc)

TABLE DES MATIERES

Conférences invitées

Reading: Earlier and more is better, so what can the NLP community do about it?.....1
V. CAVALLI-SFORZA

L'informatisation de l'amazighe : bilan et perspectives……..……3
F. ATAA ALLAH

L'intelligence sémantique au service de l'intelligence économique dans le contexte des Humanités numériques et des données massives……....5
H. FADILI

Contributions

Automatic Amazigh text readability metric: An exploratory study……..7
S. BOULAKNADEL, F. ATAA ALLAH

Relevant features selection in readability assessment……15
N. NASSIRI, V. CAVALLI-SFORZA, A. LAKHOUAJA

Recommandation personnalisée des ressources pédagogiques dans un MOOC par combinaison de filtrage collaboratif et d'analyse sémantique du contenu……25
Y. CHAABI, N.M. NDIAYE, K. LEKDIOUI

Paramètres optimaux pour la reconnaissance automatique des chiffres et lettres amazighs à base de Hidden Markov Model Toolkit……41
S. EL OUAHABI, M. ATOUNTI, M. BELLOUKI

Free large-scale Arabic speech corpus for speech synthesis…….55
O. ZINE, A. MEZIANE

Phonemes recognition using formant analysis in the case of consonant vowel transition case "Amazigh Language"……67
A. OUHNINI, B. AKSSASE, M. OUANAN

Dictionnaire général de la langue amazighe informatisé……...79
F. ATAA ALLAH, J. FRAIN, S. BOULAKNADEL

Feature selection from an expanded Moroccan opinionated dataset…….91
I. TOUAHRI, A. MAZROUI

La non voyelle dans la langue arabe……99
I. HADJIR, F.Z. BELKREDIM, A. ABBACHE, A. GHOULAM

Chunker de la langue arabe basé sur les règles de la grammaire classique……105
N. ABABOU, A. MAZROUI, R. BELEHBIB

Tifinagh handwritten character recognition using deep learning…….119
D. EL WARDANI

Using a novel approach for automatic concepts extraction from Arabic texts……..125
A. GHOULAM, A. ABBACHE, F.Z. BELKREDIM, I. HADJIR

Injecting linguistic features into Arabic-English factored translation model…….……....133
S. BERRICHI, A. MAZROUI

Une nouvelle approche d'analyse sémantique des tweets pour l'extraction des traits de personnalités……..…143
M. ECHHAIMI, Y. CHAABI, T. BOUJIHA

Reading: Earlier and more is better, so what can the NLP community do about it?

Violetta CAVALLI-SFORZA

School of Science and Engineering, University Al Akhawayn, Ifrane, Morocco

Abstract.

The invention of writing and subsequent evolution of different alphabets marked a revolution in the socio-cultural history of humanity. With technology able to record history, creative literature and accumulated knowledge, reading became an essential skill for people to gain access to that knowledge in order to further their socio-economic position, as well as to entertain themselves and be inspired to creatively contribute to the written record. Notwithstanding the long-accepted value of being able to read, many parts of the world—both "developed" and "developing"—still have significant pockets of adult illiteracy, partial or complete, in spite of the many efforts to combat the situation; even among school children, reading skills are sometimes sadly lacking. The ability to read, like that of learning a new language, is most easily acquired early in life, when the human brain is at its most plastic. Failure to acquire the expected reading level by 3rd grade has significant impact on academic achievement because, to this day, reading is how we mostly acquire academic knowledge. Stepping further back, research into early childhood cognitive development increasingly shows that the level of reading and language skills attained by children in the early primary school grades is strongly correlated to the exposure they had to language as infants, and to whether they were read to at home and how much. These opportunities are themselves heavily correlated with socio-economic status. In light of the current literacy situation of Morocco and other countries, and the recent scientific findings about when pre-literacy really begins to develop, we should ask ourselves as language technology researchers and responsible human beings whether and how our profession can assist in creating and distributing resources that will enhance literacy skills from a very young age.

Biography.

Violetta Cavalli-Sforza is Associate Professor of Computer Science at Al Akhawayn University, where she has taught since 2008. She has a B.S. and a M.S. in Civil Engineering (1980) and a M.S. in Computer Science-Computer Engineering (1985) from Stanford University. She obtained her Ph.D. in Intelligent Systems from the University of Pittsburgh in 1998, working on computer-assisted instruction. She worked as staff member (1994-1999) and visiting researcher (2003-2007) at Carnegie Mellon University on machine translation projects, and as an assistant professor at San Francisco State University (2000-2002). She has also consulted for the Linguistic Data Consortium and other organizations on Arabic-related projects (2003-2008, 2012, 2013-2015). Her interest in the Arabic language, first as a language learner and then for Natural Language Processing, led to research funding from the National Science Foundation and Fulbright for working on Arabic in Morocco, where she eventually settled. Her teaching interests include natural language processing, artificial intelligence, languages and compilers, and data structures; she has also promoted the development of undergraduate research courses and multidisciplinary courses at Al Akhawayn University. Her research interests focus on: artificial intelligence applied to computer-assisted learning in general and language learning specifically, especially for reading and more recently for writing; resource construction for the Arabic and Amazigh (Berber) languages and their regional variants; and visualization of information.

L'informatisation de l'amazighe : bilan et perspectives

Fadoua ATAA ALLAH

CEISIC, Institut Royal de la Culture Amazighe, Rabat, Maroc

Résumé.

Durant ces dernières décennies, entrainées par l'évolution permanente de l'informatique et la constante informatisation de nos sociétés, les technologies du langage ont connu des progrès majeurs. Le déploiement de ces technologies pour des langues peu-dotées représente un grand défi, notamment que les technologies du langage sont pleinement confrontées au manque des ressources linguistiques et représentent ainsi peu la diversité des langues. Comme beaucoup d'autres, la langue amazighe est aussi touchée par cette lacune.

Dans cette communication, il est question de la stratégie adoptée en matière de technologies numériques de l'Institut Royal de la Culture Amazighe (IRCAM) pour préserver la langue et la culture amazighes afin de permettre de répondre aux besoins d'expression et de communication des citoyens et des institutions.

La stratégie de l'IRCAM en matière de l'informatisation consiste à orienter son action autour de deux priorités : prendre en compte la dimension linguistique des technologies numériques et contribuer à mettre celles-ci au service de l'amazighe. Néanmoins, la voie de la concrétisation de cette stratégie est pleine de défis et d'entraves : elle nécessite, en effet, la coordination des acteurs concernés par le développement et la transmission de l'amazighe dans les domaines de la politique linguistique, la recherche scientifique et la communauté du logiciel libre.

Biographie.

Fadoua Ataa Allah a obtenu un doctorat en Sciences de l'Ingénieur, spécialité Informatique et Télécommunication à la Faculté des Sciences de Rabat, Université Mohamed V - Agdal en collaboration avec l'Université de Michigan-Dearborn aux Etats-Unis. Elle a travaillé, depuis 2008, au sein du Centre des Etudes Informatiques, des Systèmes d'Information et de Communications, à l'Institut Royal de la Culture Amazighe, en tant que chercheur spécialisée en traitement automatique des langues.

De 2011 à 2015, elle a occupé le poste de coordination de l'unité d'étude et de recherche du traitement de l'information et de communication ; et depuis 2016, elle est directrice dudit centre. Elle a piloté et réalisé plusieurs projets en traitement automatique de la langue amazighe, notamment ceux liés à la morphologie amazighe, les ressources langagières et la traduction automatique. Elle a aussi largement contribué à l'encadrement de jeunes chercheurs dans différentes thématiques. En outre, elle est l'auteur ou le coauteur de nombreuses publications nationales et internationales.

L'intelligence sémantique au service de l'intelligence

Hammou FADILI

CNAM, Paris, France

Résumé.

Les travaux et expériences de recherches qui sont présentés s'inscrivent dans une stratégie générale de recherche sur les humanités numériques et sur l'intelligence économique, en particulier, et dans leur orientation relative à l'exploitation des grandes masses de données.

Dans ce contexte, les données proviennent principalement des avis et des opinions des utilisateurs/clients, relatés via des réseaux sociaux, des forums, des blogs, des sites de ventes, etc., sur toutes sortes de sujets tels que les événements, les produits, les attitudes, etc., pour exprimer leurs sentiments et leurs expériences. Les phénomènes d'analyse des données et en particulier économiques et sur les humanités dans le contexte du « Big data », pour les besoins de renseignement, questionnent les méthodes et démarches traditionnelles de modélisation, de conception, d'implémentation et d'évaluation des systèmes d'information et de décision.

Biography.

Hammou Fadili is a researcher HDR (Habilité à Diriger des Recherches) at Conservatoire National des Arts et Métiers (CNAM) in computer science. He is a member of the information systems team, named Ingénierie des Systèmes d'Information et de Décision (ISID). Also, he is a manager of the Digital Humanities projects at the research department of the FMSH and associate researcher at the Digital Humanities department of the Paris 8 University. His research work concerns especially these areas:

- Semantic Web, WEB 3.0,
- Machine learning (supervised, unsupervised, etc.),

- Digital Humanities,
- Automatic Natural Language Processing (NLP),
- Language & Context Modeling,
- Detection and anonymisation of sensitive data,
- Semantic mining of structured and unstructured data,
- (Linked, Big, Smart) data.

He is a member of the program committee and member of the organization committee for various conferences. He is also reviewer for some journals. He is (or was) involved in many ANR or European research projects.

DiLiTAL'19, p. 7-14

Automatic Amazigh text readability metric: an exploratory study

Siham BOULAKNADEL, Fadoua ATAA ALLAH

Center of Computer Sciences Studies, Information Systems and Communication, Royal Institute of Amazigh Culture

{boulaknadel,ataaallah}@ircam.ma

Abstract.

There has been a lot of research in the field of text readability for English, French, Spanish, and Arabic languages. To the best of our knowledge, there has not many researches on automatic Amazigh readability measurements. This paper presents our approach to develop a prototype tool to automate the task of measuring Amazigh text readability using factor analysis. Our results are based on Amazigh texts extracted from the manuals for the elementary classes (second grade through sixth grade). Preliminary evaluation results along with future directions of such research are presented.

1. Introduction

Text readability can be expressed as the ease of understanding or assimilation a given text. It is due to many associated factors, such as: the style of writing, the length of the sentences, the proportion of difficult words contained in the text, etc. Readability levels can help educators in selecting the appropriate text for their audience and evaluating the desired impact on its readers. Readability has some potential application areas such as education in order to select the appropriate text for their audience, business publications, complex financial reports (Yan *et al*., 2014), online media (Leong *et al*., 2002), and health care (Berland *et al*., 2001).

Many approaches have used traditional or modern methods to achieve a valid readability metric. Traditional methods are based on using mathematical formulas, for instance: The Flesch Reading Ease, Flesch-Kincaid Grade Level, SMOG Index, Gunning Fog Index, Automated Readability Index and Dale-Chall readability formula. Modern methods are based on machine learning models, which is scalable to a large number of

features and can accept features generated automatically from electronic documents.

The remainder of the paper is structured as follows. Section II gives a brief review of related work. Section 3 introduces the main features of Amazigh language. Section 4 presents the data used for the experiment. Section 5 presents the proposed processing for calculating the Amazigh text readability and the evaluation results. Finally, Section 6 gives conclusions and directions for future works.

2. Related works

Developing reading ability is an essential part of language acquisition. However, finding proper reading materials for a learner is not simple. Focusing on that, the first studies were begun with Lucius Adelno Sherman, in 1880 (Du Bay, 2004). He points up to the importance of average sentence length and established that shorter sentences boost readability.
Then, they were followed by the readability evaluation, where many measures' formulas were proposed for English (Flesch, 1943), Polish (Pisarek, 1969), French (Conquet and Richaudeau, 1973). Within several years, dozens of readability formulas, based on text statistics, were proposed. However, the five common are presented in the following table (Senter and Smith, 1967; Gunning, 1968; McLaughlin, 1969; Coleman and Liau, 1975; Kincaid *et al*., 1975; Zhou *et al.*, 2017).
Lately, readability measures have been employed for measuring the reading grade level of the textbooks and the validity of the formulas have been evaluated not only for English but also for many different languages.

Besides these classical methods to examine readability, the modern studies use machine learning approaches, such as SVM and Naïve Bayes classifiers to estimate the grade level of a given text (Schwarm and Ostendorf, 2005). Most of these approaches combine traditional reading metrics with statistical language models, based on textual features related to lexical, syntactic, and semantic levels (Tanaka-Ishii *et al.*, 2010; El-Haj and Rayson, 2016).

Readability Formula	Formula Content	Shallow Features
Flech-Kincaid Grade Level	$0.39\left(\frac{total\ words}{total\ sentences}\right) + 11.8\left(\frac{total\ syllables}{total\ words}\right.$	Average Sentence Length, Syllables-based
SMOG	$1.0430\sqrt{hard\ words \times \frac{30}{sentences}} + 3.1291$	Hard words (more than three syllables)
Gunning Fog	$0.4 \times [\frac{words}{sentences} + (100 \times \frac{hard\ words}{words})]$	Average Sentence Length, Hard words
Coleman-Liau	0.0588x(average number of letters per 100 words) - 0.296x(average number of sentences per 100 words) - 15.8	Average number of letters, average number of sentences
Automated Readability	$4.71\left(\frac{letters}{words}\right) + 0.5\left(\frac{words}{sentences}\right) - 21.43$	Average Sentence Length, Character-based

Table 1. Five common readability measures

3. Amazigh language

Amazigh is a language from the Hamito-Semitic branch (Greenberg, 1966). It is currently presented in a dozen countries ranging from Morocco, with 50% of the overall population, to Egypt, passing through Algeria with 25%, Tunisia, Mauritania, Libya, Niger and Mali. In Morocco, we distinguish three major Amazigh dialects: Tamazight of the north, known as Tarifit, Tamazight in the center, and Tashelhit in the south.
Its graphic system "Tifinagh", written from left to right, contains 33 alphabets (Ameur *et al.*, 2004). Morphologically, Amazigh is a non-concatenative language (Ataa Allah, 2014). It is a highly inflected and also shows derivation to high degree. Syntactically, the canonical order of the Amazigh sentence is basically VSO (Verb, Subject, direct Object). That makes the Amazigh processing more difficult.

However, there has been an increasing attention in Amazigh language processing. Despite this growing interest, there has not many researches that have addressed the problem of measuring a proper Amazigh readability metric (Boulaknadel and Ataa Allah, 2016). In this context, our research attempts to advance the research in Amazigh readability and to computerize the process of measuring Amazigh text readability in a manner similar to other languages.

4. Data experiment

The used corpus contains text extracted from the Amazigh curriculum for all elementary classes (second grade through sixth grade) (Boudriss *et al.*, 2009; Boudriss *et al*., 2005; Boudriss *et al*., 2006; Agnaou *et al*., 2007; Agnaou *et al*., 2008). Since some of the Amazigh curriculum is not available in an electronic format, we resorted to entering some collection of Amazigh text manually. The collection contains more than 78101words.
Before analyzing the Amazigh textbooks, we have performed the following pre-processing steps:

- Normalization, which consists of the conversion of the character encoding to UTF-8 Unicode.
- Sentence tokenization, which is the segmentation of document content into sentences.
- Word tokenization that is the process of breaking up a text into words. In this step, multiple whitespace characters are replaced by a single whitespace character.

5. Proposed processing

In this section, we describe our approach in extracting features from our Amazigh annotated corpus for readability level.

Measuring readability is the process of extracting all possible features of a text and supplies them with representative numerical value(s). There are many factors related with text's lexical content and are considered important to determine the readability of any given text. In this step, we extracted seven features that have been employed in numerous readability formulas.

1. Number of characters in a text.
2. Number of words in a text.
3. Number of sentences in a text.
4. Number of difficult words in a text, where our definition of a difficult word is the words consisting of more than six letters.
5. Average sentence length in a text.
6. Average word length in a text.
7. Average number of difficult words in a text.

In order to evaluate our corpus, we used, in our experiments, several readability metrics. These metrics are Laesbarheds-Index (LIX) (Björnsson, 1968), Automated Readability Index (ARI) (Si *et al*., 2001), SMOG (McLaughlin, 1969), and Gunning Fog (Gunning, 1968).

These formulas were selected for their simplicity and more notably due to their parameters that can be easily applied to Amazigh texts. They are also selected for the fact that they do not use language-dependent features like the number of syllables in a word, as it is the case in Gunning-Fog metric. ARI readability formula score refers to the US grade level needed to understand the text. LIX readability formula gives a score that determines the difficulty of a given text. Table 2 lists LIX readability scores and their meanings.

Scores	Meaning
0-24	Very easy
25-43	Easy
35-44	Standard
45-54	Difficult
55 and above	Very difficult

Table 2. LIX readability scores and their meanings

SMOG determines the reading level of written material. It focuses on the length of words and sentences rather than on words alone.
Whereas, the Gunning-Fog formula is a weighted average of the number of words per sentence, and the number of long words per word. It translates the number of education years a reader needs to understand the material. The ideal score is 7 or 8; anything above 12 is too hard for most people to read.

We have computed the seven main features previously mentioned and calculated the readability formulas for our collection of Amazigh texts. The results of our statistical measures are summarized in Tables 3 and 4.

In Table 2, we notice that the results are not as simple to take to mean. There is no noticeable trend in both Number of words and Number of sentences. Whereas, the presence of trends is important to determine the most important parameters that affect and represent the text's Amazigh readability.

Grade Level	Number of Characters	Number of Words	Number of Sentences
2nd Grade	67 682	17 464	2 344
3rd Grade	50 135	13 221	1 803
4th Grade	51 447	13 119	2 001
5th Grade	55 621	13 968	2 325
6th Grade	73 102	18 717	2 902

Table 3. Text statistics for Amazigh text collection

We have also computed LIX, ARI, SMOG and Gunning-Fog readability formula values for all the text grades as shown in Table 3. As we can constate, only LIX score gives close to steady results in representing the readability of an Amazigh text. As LIX was not proposed to be employed with Amazigh texts, the LIX results evaluate all the processed Amazigh texts, corresponding to the five grades, as very-easy.

Both ARI and Gunning-Fog provide incoherent results. Moreover, we have noticed that ARI gives intolerable negative score values for some Amazigh texts.

Grade Level	LIX Score	ARI Grade Level	SMOG	Gunning Fog
2nd Grade	19,0686979	0,54893325	7,62747917	8,44414757
3rd Grade	21,7719285	0,0970542	8,70877142	9,00737686
4th Grade	20,1700644	0,31867806	8,06802576	8,52622055
5th Grade	20,0899298	0,32923394	8,03597192	8,38365131
6th Grade	19,8172167	0,19044359	7,92688669	8,43354419

Table 4. Average readability scores for Amazigh text collection

Due to the ambiguity of the correlation between the extracted text lexical features and the processed Amazigh text, in addition to the inexactness and inapplicability of other tested readability measures on Amazigh texts, the need is emphasized to consider a new Amazigh readability measure.

6. Conclusion

In this paper, we conducted an analysis based on our collection of Amazigh texts, with more than 78101 words extracted from the Amazigh curriculum, and a set of metrics, selected from the most common used measures for evaluating readability. We confirmed the drawbacks of the existing readability formulas, and their ineffectiveness to provide truthful readability scores for Amazigh texts.

In the future work, we plan to consider an Amazigh readability formula, which take into account the most influential text features.

References

Agnaou F., El Baghdadi M., Laaraj H., Demntati M., Khalafi A., Moukhlis M., Eazzi E.M., Souifi H., Kassi A., Manuel scolaire de la cinquième année du primaire, « tifawin a tamazityt 5 » (ⵜⴼⴰⵡⵉⵏ ⴰ ⵜⴰⵎⴰⵣⵉⵖⵜ 5), livre de l'élève, Imprimerie ImpriMahd, Casablanca, 2007.

Agnaou F., El Baghdadi M., Laaraj H., Demntati M., Khalafi A., Amrous N., Eazzi E.M., Souifi H., Kassi A., Manuel scolaire de la sixième année du primaire, « tifawin a tamaziɣt 6 » (ⵜⴼⴰⵡⵉⵏ ⴰ ⵜⴰⵎⴰⵣⵉⵖⵜ 6), livre de l'élève, Imprimerie ImpriMahd, Casablanca, (2008).

Ameur M., Bouhjar A., Boukhris F., Boukouss A., Boumalk A., Elmedlaoui M., Iazzi E. M., Souifi H. (2004). *Initiation à la Langue Amazighe*. IRCAM, Rabat.

Ataa Allah F. (2014). Finite-State Transducer for Amazigh Verbal Morphology. *Literary & Linguistic Computing*, Oxford University Press.

Berland G. K., Elliott M. N., Morales L. S., Algazy J. I., Kravitz R. L., Broder M. S., Kanouse D. E., Muñoz, Puyol J., Lara M., Watkins K. E., Yang H., McGlynn E. A. (2001). Health information on the Internet: accessibility, quality, and readability *in* English and Spanish. *Jama* 285(20): 2612-2621.

Boudriss B., Agnaou F., El Baghdadi M., Eazzi E.M., Demntati M., Khalafi A., Souifi H., Lhbib F., Manuel scolaire de la deuxième année du primaire, « tifawin a tamazityt 2 » (ⵜⴼⴰⵡⵉⵏ ⴰ ⵜⴰⵎⴰⵣⵉⵖⵜ 2), livre de l'élève, Edition Alouma, Casablanca, 2004.

Boudriss B., Agnaou F., El Baghdadi M., Eazzi E.M., Demntati M., Khalafi A., Souifi H., Lhbib F., Manuel scolaire de la troisième année du primaire, « tifawin a tamazityt 3 » (ⵜⴼⴰⵡⵉⵏ ⴰ ⵜⴰⵎⴰⵣⵉⵖⵜ 3), livre de l'élève, Imprimerie OKAD, Rabat, 2005.

Boudriss B., Agnaou F., ElBaghdadi M., Eazzi E.M., Demntati M., Khalafi A., Souifi H., Lhbib F., Manuel scolaire de la quatrième année du primaire, « tifawin a tamazityt 4 » (ⵜⴼⴰⵡⵉⵏ ⴰ ⵜⴰⵎⴰⵣⵉⵖⵜ 4), livre de l'élève, Imprimerie OKAD, Rabat, 2006.

Boulaknadel S., Ataa Allah F. (2016). Vers un modèle de prédiction de la difficulté lexicale. *Proceeding of a conference on 'De la langue maternelle à la langue de l'école'*, Rabat, Morocco, 14-15 november.

Coleman M., Liau T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283-284.

Conquet A., Richaudeau F. (1973). Cinq méthodes de mesure de la lisibilité. *Communication et Langages*, Vol. 17, pp. 5-16.

Du Bay W. H. (2004). *The Principles of Readability*. Impact Information.

El-Haj M., Rayson P. E. (2016). OSMAN: a novel Arabic readability metric. *Proceedings of the Language Resources and Evaluation Conference*. Portoroz, Slovenia, pp. 250-255.

Flesch R. F. (1943). Marks of Readable Style: A Study in Adult Education. Number 897 in Contributions to education. Teachers College, Columbia University.

Greenberg J. H. (1966). *The Languages of Africa*, 2nd ed. Mouton, The Hague.

Gunning R. (1968). *The Technique of Clear Writing, revised ed*. New York, USA: McGraw-Hill.

Kincaid J. P., Fishburne R. P., Rogers R. L., Chissom B. S. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease formula) for Navy enlisted personnel. *Naval Air Station Memphis*, Millington, TN, USA.

Leong E. M. T., Ewing M. T., Pitt L. F. (2002). E-comprehension Evaluating B2B websites using readability formulae. *Industrial 2 Marketing Management*, 31:125-131.

Mc Laughlin G. H. (1969). SMOG grading: A new readability formula. *J.* Reading, 12(8):639-646.

Pisarek W. (1969). Jak mierzyé zrozumiałość tekstu. *Zeszyty Prasoznawcze*, 4(42):35-48.

Schwarm S. E., M. Ostendorf. (2005). Reading level assessment using support vector machines and statistical language models. *Annual Meeting of the ACL*, pp. 523-530, Ann Arbor, Michigan.

Senter R., Smith E. (1967). Automated readability index. Aerospace Medical Research Laboratories, US Air Force, Wright-Patterson Air Force Base, OH, USA.

Tanaka-Ishii K., Tezuka S., Terada H. (2010). Sorting texts by readability. Computational *Linguistics*, 36(2):203–227.

Yan S., Wang X., Yu Y. (2014). Readability to Financial Report: A Comparative Study of Chinese and Foreign Countries. *Computational Sciences and Optimization, Seventh International Joint Conference on. IEEE*.

Zhou S., Jeong H., Green P. A. (2017). How Consistent Are the Best-Known Readability Equations in Estimating the Readability of Design Standards?, *IEEE Transactions on Professional Communication*, 60(1):97-111.

DiLiTAL'19, p. 15-23

Relevant features selection in readability assessment

Naoual NASSIRI[1], Violetta CAVALLI-SFORZA[2], Abdelhak LAKHOUAJA[1]

[1]Department of Computer Science, Faculty of Sciences, University Mohamed First, Oujda, Morocco

[2]School of Science and Engineering, Akhawayn University, Ifrane, Morocco

{naoual.nassiri, abdel.lakh}@gmail.com. v.cavallisforza@aui.ma

Abstract.

The diversity and the growth of available data contribute to the progress of several Natural Language Processing (NLP) fields, including text readability assessment. Although in recent years, the Arabic language has significantly increased its resources, particularly in terms of corpora and syntactic tools, there are still types of resources that are not widely available, especially when it comes to corpora annotated by difficulty levels. Learning Arabic as a foreign language presents a great challenge in terms of choosing materials to be presented to the learner. In order to develop powerful tools that automate this choice, we must first choose the optimal features set to be used in the development of such a tool. In this paper, we present an approach to automatically measuring the readability of Arabic as a foreign language (L2) using a wide range of features in a first experiment and then reducing them in two subsequent experiments by ignoring irrelevant features in the readability prediction process. We obtain a best accuracy result for L2 readability prediction model of 86.04% in the validation phase with the final reduced feature set.

1. Introduction

As the amount of available data continues to increase, there is a growing need for tools to help people, especially educators, to better find, filter and manage the resources they need to present to learners. The readability of texts, i.e. the assignment of documents to a level of difficulty among several predefined levels based on their content, is an important task in education. If

done manually, this process is very long and costly, which limits its applicability. As a result, there is a growing interest in the development of technologies for the automatic measurement of text readability. A number of statistical classification and Machine learning (ML) techniques have been applied to text readability measurement, including regression models, nearest neighbor classifiers, decision trees, Bayesian classifiers, rule learning algorithms and neural networks. These techniques were first developed in the form of a mathematical equation, called "classical formulas", which were relatively simple and whose independent variables were based on counts, words, syllables or letters such as Flesch (1948) and Fog (Gunning, 1969) formulas. More recently, various techniques from NLP and ML have changed and strengthened the traditional approach. These new approaches use language models, syntax analyzers, various measures of text coherence and structure, or lexical unit analysis, among others. The majority of these studies focuses on English as a first language (L1) and believes that the application of NLP techniques in readability provides better results than the traditional approaches. With regard to Arabic as a foreign language, it seems even more necessary to adapt the features that the L1 and L2 reading processes use in other languages. It should be stressed that the L2 reader be more sensitive to difficulties in terms of vocabulary and syntax and that one or more pre-existing languages influence the reader ability. The combination of NLP and ML techniques in readability prediction requires three main steps: the first step consists of gathering a reference corpus, the second step requires identifying a set of linguistic features, also known as predictors, to use with the ML techniques. These must have a significant correlation or even a causal relationship, with the difficulty of the texts. The last step consists of a classification phase.

In our study, firstly, we considered predicting readability with a broad set of specific features for the Arabic language. Then, we tried to ignore some features that we consider irrelevant to achieve comparable or better results with a more pared down feature set. The rest of this paper is organized as follows. In section 2, we present some existing readability measurement works for Arabic as a second language, focusing mainly on the features used in each of them. In Section3, we describe the data and the tools we used. The implemented approach and details on the list of features are described in section 4. In section 5, we explore the results of the different sets of features obtained. Finally, we conclude with some thoughts on future work.

2. Related work

There has been a great deal of research on readability in European languages, but relatively little attention has been paid to the Arabic language.

This section highlights recent studies that have used ML approaches on Modern Standard Arabic (MSA) corpora for second language learning.

Most of these works are developed based on features that are commonly used in Arabic L1, Arabic L2 and even in other languages such as the number of sentences, words, characters etc., and they used morphological features that are extracted using morphological analyzers.

- **Matching an Arabic Text to a Learners' Curriculum** (Cavalli-Sforza and Mezouar, 2014): In this paper, the authors present an exploratory study aimed at constructing a predictive model to automatically assess whether a given text is suitable for a learner studying at the intermediate level. The approach is based on the extraction of some features already used for other languages, and to determine if they are suitable for measuring the readability of the Arabic language. The resulting model is poor in some cases, when it comes to predicting the exact level of the program for which a text should be used.

- **Automatic Readability Detection for Modern Standard Arabic** (Forsyth, 2014): The author describes a system that uses ML to automatically predict the difficulty level of MSA text. They extracted lexical and discursive features for each document in the corpus.

- **Text readability for Arabic as a foreign language** (Saddiki *et al*., 2015): In this study, the authors assess the usefulness of lexical, morphological and semantic features to determine the readability of MSA text. They gathered 35 features of low complexity in order to establish a baseline for future research on readability assessment. The results suggest that a small set of easily computable features can be indicative of the reading level of a text.

- **Modern Standard Arabic Readability Prediction** (Nassiri *et al.*, 2018b): In this work, the authors present an approach to predicting the readability of MSA educational texts. This method is based on 170 text features dedicated to Arabic. The results obtained are very encouraging and better than the state of art.

- **Arabic Readability Assessment for Foreign Language Learners** (Nassiri *et al*., 2018a): This work is a follow-up of the previous work. The authors present an approach based on 133 text features. This new set of features was obtained by changing the morphosyntactic analyzers used, which subsequently allowed the combination of some features. The results obtained are better than those obtained in previous work.

3. Methodology

This section describes the resources used to build and evaluate a readability prediction model for L2 MSA texts. Our resources include a corpus of documents annotated by readability level, an MSA frequency dictionary, morphosyntactic analyzers and a supervised ML system. We describe each of them below.

3.1. Tools

In this study, we use morphosyntactic analyzers to annotate the data and a ML tool for the classification step.

- **Alkhalil-Toolkit** (Ababou and Mazroui, 2016; Boudchiche and Mazroui, 2018): Alkhalil-toolkit provides full morphological disambiguation, part-of-speech tagging, and other useful information. Alkhalil's part of speech tagger specifies, for each word, the proclitics, the lemma, and the enclitics. Alkhalil-toolkit includes other features such as stems, and roots. The toolkit analyzers are all based on Alkhalil Morpho Sys 2 (Boudchiche *et al*., 2017), they provide all their tags taking into account the context of the analyzed word.

- **WEKA** (Holmes *et al.*, 1994): The Waikato Environment for Knowledge Analysis is an open-source ML software tool, based on a set of classification algorithms. It contains data pre-processing, classification, regression, clustering, association rules, and visualization tools.

3.2. The frequency dictionary

The use of features in Arabic readability research is derived from different types of word lists. Modern frequency lists are more robust than early ones because they are typically based on multi-million word corpora. In this work, we built and used a 5000-word frequency dictionary of Arabic, based on the Tashkeela corpus (Zerrouki and Balla, 2017), which is composed of 79 million words. Each entry of the dictionary represents the lemma corresponding to several word forms appearing in the corpus.

3.3. The corpus

We collected a corpus from the Aljazeera-Learning website texts. Their site is open to the public and includes material for learning Arabic as a foreign language. Aljazeera-Learning's corpus contains texts that are based on authentic MSA materials, classified according to five levels of difficulty, which are Beginner1 (1), Beginner2 (1+), Intermediate1 (2), Intermediate2 (2+), and Advanced (3). The corpus contains a total of 347 documents and 50 968 tokens.

4. Feature engineering

Several features have been used in readability assessment by researchers in different languages. These features can also be applied to the readability of MSA. The length of words in terms of syllables, affixes, or the number of characters per word, is one of the traditional features that are applied to all languages. In (Habash, 2010), the author explains that, in Arabic morphology, multiple affixes and clitics can be attached to a lexical stem. This complex morphology must be analyzed by a reader and therefore can increase the difficulty in understanding.

Based on the lemma, part-of-speech, rank, and raw frequency provided by our frequency dictionary, we generated features for the corpus documents. Most of our readability features were derived from this dictionary. We obtained all of our features with the support of Alkhalil-toolkit's features and tokenization output. We included lexical and type-to-token ratio features in our model. We also included the PoS-based ratio features used first in a previous study with MADAMIRA, but adjusted them to match MSA PoS classes provided by Alkhalil-PoS tagger. Average sentence length, which originates from the traditional formulas, is also included. We used Alkhalil segmentation to delineate sentences. Altogether allow us to develop 133 features, grouped by categories, as shown in Table 1. We collected this list of features for use in the ML system.

5. Results and discussion

In this section, we will describe the results of the three experiments performed on the Aljazeera-Learning corpus.

5.1. Experiment 1

The first experiment consists of generating a model on an 80% of the data, chosen randomly. Then, the model obtained is validated on the remaining 20%. The content of the learning corpus is not balanced between the five levels in terms of the number of texts. In order to have enough texts for the validation phase, we decided to work with the 80%/20% data split instead of the usually used 90%/10% one. This experimentation is performed using all the 133 features list gathered (section 4) and is repeated for the five chosen algorithms and every algorithm is applied respectively on five, four (obtained from five classes by gathering level 1 and level 1+ into a single 1_1+ level) and three classes (obtained from four classes by gathering level 2+ and level 3 into a single 2+_3 level).

Category	Description	Number of Features	Example
PoS-based Frequency	Individual PoS type occurrences calculated based on the frequency dictionary	84	Frequent noun tokens to all noun tokens ratio
Type-to-Token PoS Ratio	Number of unique word types divided by the count of total tokens.	25	Noun-to-token ratio
Token & Type Frequency	Frequent lemmas divided by the count of total tokens.	10	Frequent type-to-token ratio
Type-to-Token	Lemmas to token ratio.	2	Distinct lemma to token ratio
Word Length	Average length of words	2	Token count
Vocabulary Load	Familiarity with Vocabulary	3	Frequent lemmas count
Word Class	Open class and closed class	4	Open class tokens count
Sentence Length	Average length of sentences	3	Average sentence token length
Total		**133**	

Table 1. The initial set of 133 features, distributed over 8 categories

Table 2 presents the validation results. We achieve a maximum accuracy value of 86.04% using OneR, SMO and Random Forest algorithms using a three-class classification, while we get 79.06% as maximum accuracy values respectively on five- and four-class classification.

	5 Classes	4 Classes	3 Classes
OneR	60.46%	**86.04%**	**86.04%**
J48	**79.06%**	74.41%	81.39%
IBK	53.48%	55.81%	55.81%
SMO	74.41%	79.06%	**86.04%**
RF	76.74%	83.72%	**86.04%**

Table 2. The133 features set validation results

5.2. Experiment 2

The second experiment also consists of an 80/20% model generation process. This experiment is performed using 129 features list obtained from the first 133 list by ignoring four "percent missing" PoS-based Frequency features (features that are equal to zero with all the corpus texts). The process is also repeated for the five chosen algorithms and every algorithm is applied respectively on five, four and three classes. The validation results are as shown in Table 3.

	5 Classes	**4 Classes**	**3 Classes**
OneR	**79.06%**	**86.04%**	**86.04%**
J48	**79.06%**	74.41%	81.39%
IBK	53.48%	55.81%	55.81%
SMO	74.41%	79.06%	**86.04%**
RF	**79.06%**	83.72%	83.72%

Table 3. The 129 features set validation results

In this second experiment, we also achieve a maximum accuracy value of 86.04% using OneR and SMO algorithms on three classes' classification and 79.06% with five and four classes classification.

5.3. Experiment 3

The third and last experiment was performed using 114 features list obtained from the previous ones. The ignored features are ratios of other existing features. The list of redundant features ignored in the third experiment is detailed by category in Table 4.

Category	**Number of features**
Word class features	2
Vocabulary load features	1
Type to token PoS features	8
Type to token features	2
Sentence Length features	2

Table 4. Irrelevant Features list

Finally, the third experiment results as presented in Table 5, shows that with this last feature set we keep the maximum accuracy value of 86.04% with the three algorithms like the first set (OneR, SMO and Random Forest) and on three classes' classification also.

	5 Classes	4 Classes	3 Classes
OneR	**79.06%**	**86.04%**	**86.04%**
J48	76.74%	79.06%	79.06%
IBK	55.81%	58.13%	58.13%
SMO	76.74%	81.39%	**86.04%**
RF	**79.06%**	83.72%	**86.04%**

Table 5. The 114 features set validation results

6. Conclusion

This paper presents a set of experiments that use ML techniques to predict the readability of texts dedicated to learners of Arabic as L2 using the Aljazeera-Learning corpus. The main goal of all of these experiments is to reduce the set of features used in the prediction process without affecting the maximum accuracy values obtained with the initial features set.

The first experiment consisted of generating a model using 133 features on 80% of the data, then validating this model on the 20% remaining texts. The model tests gave good results, with 86.04% accuracy using the OneR, SMO and Random Forest algorithms, which was achieved using three classes. The objective of the second experiment was to generate a model, applying the same approach as the first experiment, but ignoring four features from the first set, so we work with a 129-feature list. The models always gave good results and kept the same maximum accuracy value with the SMO and OneR algorithms. The last experiment used a 114-feature set obtained from the 129 one by ignoring some irrelevant features described above. The models obtained gave the same accuracy value as the first experiment and with the same algorithms. Therefore, these results show that the features we have removed do not really influence the classification process.

As a future work, we aim to further reduce the number of features in order to keep the smallest set that gives the best classification results. Additional possible directions of work include the development of an application to assess the readability of Arabic as a second language in other domains since here we are working on educational texts only.

References

Ababou N., Mazroui A. (2016). A Hybrid Arabic PoS Tagging for Simple and Compound Morphosyntactic Tags. International Journal of Speech Technology, 19(2):289-302.

Boudchiche M., Mazroui A. (2018). A Hybrid Approach for Arabic Lemmatization. International Journal of Speech Technology, pp. 1-11.

Boudchiche M., Mazroui A., Ould Abdallahi OuldBebah M., Lakhouaja A., Boudlal A. (2017). AlKhalil Morpho Sys 2: A Robust Arabic Morpho-Syntactic Analyzer. Journal of King Saud University - Computer and Information Sciences, 29(2):141-146.

Cavalli-Sforza V., El Mezouar M. (2014). Matching an Arabic Text to a Learners' Curriculum. Proceding of the 5th International Conference on Arabic Language Processing (CITALA), Oujda, Morocco.

Flesch R. (1948). A New Readability Yardstick. Journal of Applied Psychology, 32(3):221-233.

Forsyth J. (2014). Automatic Readability Detection for Modern Standard Arabic. Thesis dissertation, Brigham Young University - Provo.

Gunning R. (1969). The fog index after twenty years. Journal of Business Communication, 6(2):3-13.

Habash N. (2010). Introduction to Arabic Natural Language Processing. Synthesis Lectures on Human Language Technologies.

Holmes G., Donkin A., Witten I. H. (1994). Weka: A machine learning workbench. Proceedings of the Second Australian and New Zealand Conference on Intelligent Information Systems, pp. 1305-1314.

Nassiri N., Lakhouaja A., Cavalli-Sforza V. (2018a). Arabic Readability Assessment for Foreign Language Learners. Natural Language Processing and Information Systems, pp. 480-488. Springer International Publishing.

Nassiri N., Lakhouaja A., Cavalli-Sforza. (2018b). Modern Standard Arabic Readability Prediction. Arabic Language Processing: From Theory to Practice, pp. 120-133. Springer International Publishing.

Saddiki H., Bouzoubaa K., Cavalli-Sforza V. (2015). Text readability for Arabic as a foreign language. IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA), pp. 1-8.

Zerrouki T., Balla A. (2017). Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization systems. Data in Brief, Vol. 11, pp. 147-151.

DiLiTAL'19, p. 25-39

Recommandation personnalisée des ressources pédagogiques dans un MOOC par combinaison de filtrage collaboratif et d'analyse sémantique du contenu

Youness CHAABI[1], Ndeye Massata NDIAYE[2],
Khadija LEKDIOUI[3]

[1]CEISIC, Institut Royal de la Culture Amazighe, Rabat, Maroc

[2]Université Virtuelle du Sénégal – UVS, Dakar, Sénégal

[3]LASTID, Université Ibn Tofail, Kénitra, Maroc

chaabi@ircam.ma, ndeyemassata.ndiaye@uvs.edu.sn,
lekdioui_khadija@hotmail.com

Résumé.

Dans cet article, nous présentons un système de recommandation des ressources pédagogiques basé sur les traces d'interaction des apprenants dans une plateforme d'apprentissage à distance. Aujourd'hui, tous les apprenants suivent le même parcours d'apprentissage et ont donc accès au même contenu de formation, malgré que leurs besoins en termes d'informations soient différents. L'approche proposée offre aux apprenants la possibilité d'adapter leur parcours de formation grâce à un mécanisme permettant de recommander les activités les plus pertinentes pour eux. Cette approche consiste à analyser automatiquement les traces d'interaction des apprenants. Sur la base de ces traces, notre système propose automatiquement des recommandations personnalisées (cours, sections de cours, activités, document, etc.) aux apprenants. L'originalité de cette approche réside dans le fait que l'apprenant reste maître de son itinéraire de formation, en augmentant les informations dont il dispose pour l'aider à construire son parcours de formation. Dans cet article, nous présentons les résultats de la mise en œuvre de notre système développé dans le cadre d'une expérience menée auprès de 100 étudiants inscrits sur notre plateforme de formation à distance ainsi que les perspectives de recherche.

1. Introduction

L'apprentissage en ligne est un moyen révolutionnaire permettant de dispenser un enseignement de qualité, mais flexible comparé à l'enseignement traditionnel (face à face). De nos jours, les apprenants bénéficient de divers programmes d'apprentissage en ligne. Cependant, la grande diversité d'apprenants pose de nouveaux défis, comme par exemple un ensemble unique des ressources d'apprentissage peut être fourni à tous les apprenants (Coates *et al.*, 2005 ; Downes, 2012). En fait, les apprenants pourraient avoir des intérêts variés, ils peuvent avoir différents niveaux d'expertise et ne peuvent donc pas être traités de manière uniforme. Il est donc indispensable pour le tuteur d'évaluer les préférences, le niveau de connaissances et de maîtrise afin d'adapter le contenu de la formation. La personnalisation du contenu peut être réalisée en utilisant différentes stratégies, dont les feedbacks par recommandation de ressources numériques (Bulut, 2018 ; Wan, 2018 ; Ndiyae, 2019).

Dans ce travail de recherche, nous proposons une approche permettant d'adapter l'apprentissage de chaque apprenant grâce à un mécanisme de recommandation des plus pertinentes ressources pédagogiques (Drachsler, 2015). La liste des ressources recommandées est obtenue en utilisant un ensemble de stratégies de recommandation utilisant des approches de filtrage collaboratif et à base de contenu.

Cet article est organisé comme suit : la section 2 présente un état de l'art sur les systèmes de recommandation. La section 3 propose une nouvelle approche de recommandation des ressources pédagogiques. La section 4 expose une étude de cas en vue d'illustrer cette approche de recommandation. Finalement, la section 5 présente une conclusion ainsi que quelques perspectives.

2. Etat de l'art

Les systèmes de recommandation (Recommendation System, RS) sont un outil qui propose des ressources utiles à un utilisateur particulier. Ils sont devenus des applications fondamentales dans de nombreux domaines (Du, 2019 ; Baidada, 2019 ; Ghézala, 2019 ; Lahlou, 2017) : la recherche d'informations, e- commerce, e-learning, web, et bien d'autres. Dans la littérature, nous trouvons différentes formes de recommandation selon l'objectif et les informations disponibles. Les RS sont généralement classés en trois catégories (Negre, 2018) : **le filtrage éditorial**, **le filtrage collaboratif** et **le filtrage basé sur le contenu**. Dans cette section nous passons en revue les principales approches existantes.

2.1. Filtrage éditorial

Ce type de recommandation est adoptée lorsque le système ne dispose d'aucune information sur l'apprenant. L'objectif de cette approche est d'attirer l'attention et l'œil de l'apprenant novice afin de lui procurer l'envie de parcourir une partie du cours. Pour cela, nous pouvons présenter en premier niveau les cours, les documents, les activités pédagogiques et les nouveautés les plus populaires, etc. Cette approche est adoptée par la plupart des sites de vente en ligne.

2.2. Filtrage basé sur le contenu

Ce type de filtrage (Malone *et al.*, 1987) prend en compte le contenu d'une ressource pédagogique (document, descriptif d'un cours, d'une vidéo, etc.) et les besoins des apprenants en termes d'information. Cette technique est basée sur la recommandation des items similaires à ceux préférés par l'apprenant dans le passé (Pazzani *et al.*, 2007 ; Malone *et al.*, 1987). Si nous prenons l'exemple : lorsqu'un apprenant a tendance à consulter souvent des cours portant sur le domaine de l'intelligence artificielle, le système lui proposera des recommandations liées à l'intelligence artificielle. En effet, ces cours recommandés disposent de mots-clés communs tels que : " Apprentissage machine ", " Apprentissage en profondeur " ou " système multi-agent ", etc.

Les techniques du filtrage basé sur le contenu proposent aux apprenants des ressources pédagogiques via les approches de classification automatique de données. Les approches de classification associent à chaque apprenant un classifieur représentant son profil. Le fait de classifier compare les nouveaux items avec le profil de l'apprenant et recommande ceux qui sont les plus proches (Adomavicius *et al.*, 2005). Plusieurs techniques de classification ont été adoptées par les systèmes de recommandation basés sur le contenu dont les plus utilisées sont : les réseaux de neurones, les arbres de décision, les réseaux bayésiens (Cohen 1996 ; Mladenic, 1999).

2.3. Filtrage collaboratif

Le filtrage collaboratif est basé sur l'hypothèse : le bouche-à-oreille. Par exemple, les apprenants, qui veulent suivre une formation ou lire un document, demandent à leurs amis leurs opinions. Donc, dans le filtrage collaboratif, la sélection des cours à proposer à un apprenant ne dépend plus du contenu (filtrage basé sur le contenu), mais des opinions et des évaluations faites par des apprenants ayant des préférences similaires. Ainsi, si deux apprenants A et B ont évalué un certain nombre de cours de façon similaire, il y a de fortes chances que A aime ce que B aime, et inversement. Donc les cours que A a aimés peuvent être recommandés à B et inversement. Cette approche résout un problème important rencontré par l'approche de

filtrage basé sur le contenu, qui est le traitement des contenus multimédias (Adomavicius *et al.*, 2005 ; Sarwar *et al.*, 2000). Les étapes générales d'un système de filtrage collaboratif sont les suivantes (Adomavicius *et al.*, 2005 ; Berrut *et al.*, 2003 ; Zaier, 2010) :

- proposer à l'apprenant un ensemble de contenu pédagogique pour les évaluer ;
- construire le profil de l'apprenant en se basant sur ses appréciations ;
- utiliser ce profil pour aider l'apprenant dans ses prochaines recherches d'information.

3. Approche proposée

Les environnements d'apprentissage classiques offrent les mêmes contenus aux apprenants, malgré que leurs besoins en termes d'informations soient différents. Une fois un système d'apprentissage est développé, il sera difficile, voire impossible de l'adapter selon les besoins et préférences des apprenants. L'originalité de cette approche est que l'apprenant reste maître de son parcours de formation et que le mécanisme de recommandation augmente les informations qu'il a à sa disposition pour l'aider à construire ce parcours. Ce mécanisme prendra notamment en compte les préférences des apprenants en termes de style d'apprentissage, niveau de connaissance et traces d'interaction (Ndiyae, 2019). Pour améliorer la qualité des recommandations, en tenant compte du contenu des items, un algorithme de filtrage basé sur le contenu est proposé et fusionné avec l'algorithme du filtrage collaboratif afin de prédire les évaluations manquantes de la matrice apprenant-item.

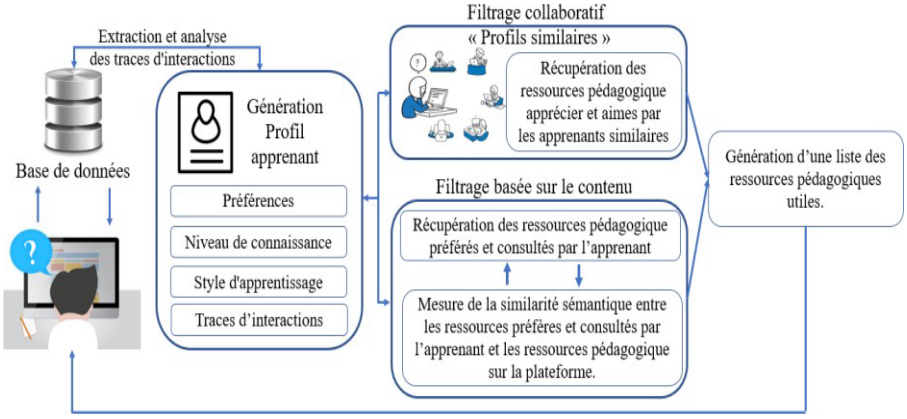


Figure 1. Architecture générale du système

Le principe général de notre approche est illustré par la figure 1. Chaque apprenant est décrit par des informations qui caractérisent son profil et il est lié par des liens d'amitié avec d'autres apprenants de la plateforme MOOC. Cette approche repose sur quatre types d'information : (1) les informations du profil apprenant qui décrivent ses connaissances, ses préférences, ses expériences, etc. ; (2) les informations qui décrivent les styles d'apprentissages de chaque apprenant ; (3) les informations qui décrivent les actions effectuées par les apprenants sur les ressources pédagogiques (les visites, les évaluations, etc.) ; (4) les informations qui décrivent les résultats des exercices ou des quiz effectués par les apprenants.

3.1. Modélisation des ressources pédagogiques

Une ressource pédagogique est décrite par un ensemble de métadonnées : titre, description, auteur, langue, format, nature, date de création, date de modification, domaine de connaissance, mots clés et difficulté. Cette dernière est utilisée dans le filtrage de ressources selon les préférences de l'apprenant liée à cette ressource. Ainsi, les ressources seront filtrées et ordonnées selon leur ordre de difficulté.

3.2. Génération du profil de l'apprenant

Tous les systèmes de recommandation ont besoin de certains types d'information sur les apprenants pour leur générer des recommandations. Les types d'information qui doivent être présentes dans un profil sont fortement liés au domaine d'applications des systèmes de recommandation. Dans notre contexte, les préférences, le style d'apprentissage et le niveau de connaissance des apprenants ont une importance cruciale.

Dans cette approche, l'apprenant doit répondre à un questionnaire pour choisir son domaine d'intérêt et le modèle du contenu qu'il préfère. Il passe deux tests pour identifier son style d'apprentissage (Vizeshfar, 2018), et un test de connaissance constitué des pré-tests de différents niveaux (faible, moyen, fort) afin d'évaluer son niveau.

3.2.1. Identification du niveau de connaissance

Dans le scénario pédagogique d'un cours en ligne, le test de connaissance constitue un dénominateur commun de la structure du cours. De manière générale, un cours en ligne est composé de 3 parties : (1) **Le système d'entrée** qui permet de gérer les flux d'apprenants à l'entrée du cours. Les objectifs généraux du cours sont présentés dans cette partie et un test d'entrée est proposé. Dans les MOOC, c'est une vidéo qui est proposée. (2) **Le système d'apprentissage** qui comprend l'ensemble des contenus et des activités d'apprentissage du cours. (3) **Le système de sortie** qui est composé

d'évaluation sommative permettant d'attester les acquis de l'apprenant. Le système d'apprentissage est composé d'unités d'apprentissage créées en fonction de l'objectif du cours. Dans chaque unité d'apprentissage est inséré un test de connaissance qui permet à l'apprenant de tester ses connaissances à la suite de son apprentissage. Le test de connaissance est constitué d'un ensemble de questions à choix multiples (QCM) et/ou de questions à choix unique (QCU). Un QCM ou QCU est souvent considérée comme étant très guidée, mais elle répond bien à la contrainte technique liée au multimédia : (i) les dispositifs de formation à distance intègrent généralement un outil de conception de QCM/QCU ; (ii) les réponses correctes sont implémentées à l'avance dans la plateforme de formation en ligne et (iii) le feedback est automatique.

3.2.2. Identification du style d'apprentissage

L'identification du style d'apprentissage des apprenants permet de varier les stratégies d'apprentissage de façon à satisfaire leurs besoins et à optimiser leur apprentissage. Le style d'apprentissages d'un apprenant représente la méthode par laquelle elle apprend le mieux, mais il peut aussi présenter une combinaison des autres styles, selon la situation. Les comportements physiologique, affectif, sociologique et cognitif représentent les indicateurs permettant d'identifier le style d'apprentissage des apprenants (O'Keefe, 1979).

Les styles d'apprentissage sont souvent classés de la manière suivante : adaptateur, divergeur, assimilateur et convergeur (voir Tableau 1).

À partir du modèle des styles d'apprentissages, nous avons conçu un questionnaire. Ce questionnaire est composé par 24 questions extraites des travaux de Felder (1996). Le temps estimé pour remplir ce questionnaire a été d'environ cinq minutes. Aucune difficulté concernant les questions posées dans le questionnaire n'a été signalée par les apprenants.

L'environnement d'apprentissage récupère les réponses et résultats de tests effectués par les apprenants et met à jour leurs profils. Les traces d'interaction des apprenants et les données sociales sont aussi utilisées dans la mise à jour du profil.

Styles d'apprentissage	Description	Recommandations
Adaptateur	Une formation doit inclure des activités de réalisation des projets, des espaces de collaboration avec pairs.	- Cours pratiques pour l'exécution et la réalisation des projets. - Espace de discussion pour la collaboration et l'échange.
Divergeur	Une formation doit inclure des périodes de discussion et d'échange entre tuteurs et apprenants, afin d'illustrer certaines notions aux apprenants et leur permettre d'échanger des idées.	- Forum de discussion et d'échange des idées entre tuteurs et apprenants.
Assimilateur	Une formation doit inclure des notions théoriques comme les études de cas.	- Cours inclure des notions théoriques comme les études de cas.
Convergeur	Une formation doit permettre la résolution de problèmes. Elle doit inclure des activités guidées par la pratique (démonstration, études de cas, etc.).	- Une formation permettant la résolution de problèmes. - Des activités guidées par la pratique (démonstration, études de cas, etc.).

Tableau 1. Les recommandations pour chaque style d'apprentissage

3.3. Similarité des profils

Dans le cadre d'apprentissage à distance, les apprenants sont intéressés à obtenir des recommandations appropriées pour leurs besoins pédagogiques. Ils peuvent être intéressés par les ressources les plus populaires et récemment visitées par leurs apprenants similaires.

Le voisinage d'un apprenant est représenté par les apprenants qui lui sont plus proches considérant les informations stockées dans le profil de chaque apprenant : les préférences, les intérêts, les styles d'apprentissage, le niveau de connaissance et les traces d'interaction. En effet, il suffit de connaître la proximité des utilisateurs en fonction de leur profil. Nous avons constaté qu'un système de recommandation efficace doit avoir les fonctionnalités suivantes :

- Construction automatique des groupes et l'affectation des apprenants aux différents groupes en se basant sur leurs profils.
- Recherche des voisins pertinents de l'apprenant actif.
- Recommandation des ressources pédagogiques appréciées par les apprenants de mêmes groupes.

Le calcul de similarité entre apprenants joue un rôle très important dans les méthodes de recommandation basées sur le voisinage (Chekkai, 2011) parce qu'il permet d'identifier les voisins de confiance. De plus, la similarité possède un impact significatif sur la précision et la performance des systèmes de recommandation (Du, 2019). Pour calculer le voisinage V entre deux apprenants x et y, nous employons la distance euclidienne.

Soient (x_1; x_2; … ; x_n) et (y_1; y_2; … ; y_n) les représentations vectorielles relatives aux profils de deux apprenants x et y. La distance euclidienne entre les deux profils est calculée comme suit :

$$V(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)} = \sqrt{\sum_{i=1}^{n} (s_i)}$$

avec $s_i \in \{0, 1\}$ tel que $s_i = 0$ si $x_i = y_i$ et $s_i = 1$ si $x_i \neq y_i$

Dans notre contexte, nous considérons uniquement trois attributs (n=3) qui sont le test de connaissance, le style d'apprentissage et préférences.

Le système de recommandation cherche les profils similaires à l'apprenant courant, par la suite il propose une liste des ressources utiles aimées et appréciées par les mêmes apprenants du groupe. La liste des ressources r recommandées à l'apprenant p est déterminée par la formule suivante (Guy *et al.*, 2010) :

$$\sum_{v \in Sim[App]}^{RV(p,r)} Ress^{-\alpha t(v,r)}$$

avec :

- App : les apprenants du système.
- RV (p, r) : les ressources r qui sont visualisées par un apprenant p. La fonction RV est égale à 1 si l'apprenant p a consulté la ressource r et 0 dans le cas contraire.
- Sim[App] : l'ensemble des apprenants qui ont un profil similaire à l'apprenant p.
- t(v,r) : le nombre de jours depuis la date de la dernière consultation de la ressource r par l'apprenant p.
- α étant un facteur de décroissance. Plus un apprenant a consulté une ressource récemment, plus le score de cette dernière augmente.

Un apprenant peut également se voir recommander une liste de ressources votées et évaluées par ses apprenants du groupe et qui sont utiles par rapport

aux domaines d'apprentissage de cet apprenant. La liste des ressources recommandées à l'apprenant p est déterminée par la formule :

$$u: P \times R \rightarrow \mathbb{R}$$

avec P l'ensemble de tous les apprenants, R l'ensemble des ressources pédagogiques possibles qui peuvent être recommandés, et u la fonction qui mesure l'utilité d'une ressource r à l'apprenant p.

Alors pour chaque apprenant p ∈ P, nous voulons choisir la ressource r′ ∈ *R* qui maximise l'utilité à l'apprenant :

$$\forall r \in R, i'_r = argmax_{r \in R} u(p, r)$$

3.4. Recommandation basée sur le contenu

La recommandation basée sur le contenu consiste à analyser le contenu des items candidats à la recommandation ou les descriptions de ces items. Dans cette phase, notre système récupère le niveau de connaissance de l'apprenant. En fonction de ce niveau, le système mesure la similarité sémantique entre les préférences, le contenu et les descriptions des ressources déjà préférés par cet apprenant dans le passé avec les ressources présentes sur la plateforme.

La similarité sémantique ou la relation sémantique est un concept de mesure de la proximité entre des termes ou documents dans le contexte de leur signification. Nous avons deux méthodes différentes pour calculer la similarité sémantique. L'une consiste à définir une similarité topologique, en utilisant l'ontologie pour définir une distance entre les mots (Mestrovic, 2016). L'autre est basée sur l'utilisation des moyens statistiques tels que le modèle d'espace vectoriel pour établir une corrélation entre les mots et les contextes textuels à partir d'un corpus de texte approprié. Nous nous concentrons sur la première approche utilisant l'ontologie WordNet pour le calcul de similarité sémantique. Le calcul de similarité dans cette approche repose sur le fait que la similarité dépend des caractéristiques communes et distinctes des objets.

WordNet est une ontologie lexicale. Il s'agit d'un réseau sémantique développé par l'Université Princeton qui modélise le savoir lexical dans une hiérarchie taxonomique. WordNet contient trois bases de données : une pour les noms, une pour les verbes et l'autre pour les adverbes et les adjectifs (Sudha, 2016). La similarité sémantique dans WordNet peut être calculée par la méthode de la longueur du chemin. Cette méthode permet de calculer le nombre des nœuds ou de relations entre les nœuds dans la taxonomie. L'avantage de cette méthode est qu'elle n'est tributaire ni de la répartition statique du corpus ni de la distribution des mots.

Nous utilisons WordNet 2.1 qui contient neuf hiérarchies de noms distinctes, où parfois le chemin entre deux concepts peut ne pas exister (voir figure 2). Par conséquent, nous créons un nœud racine "Entité" (voir figure 2) qui reprend toutes les neuf hiérarchies données dans WordNet.

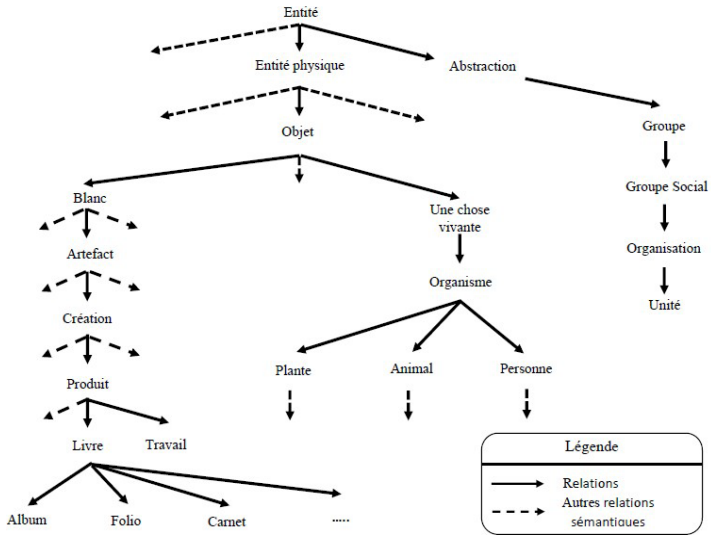


Figure 2. Extrait de la hiérarchie nominale de WordNet

Le processus de calcul de similarité sémantique se compose de trois phases (voir figure 3) :

– Phase 1: Module de construction temporaire.
– Phase 2: Module de calcul sémantique.
– Phase 3: Procédures de mesure de similarité sémantique.

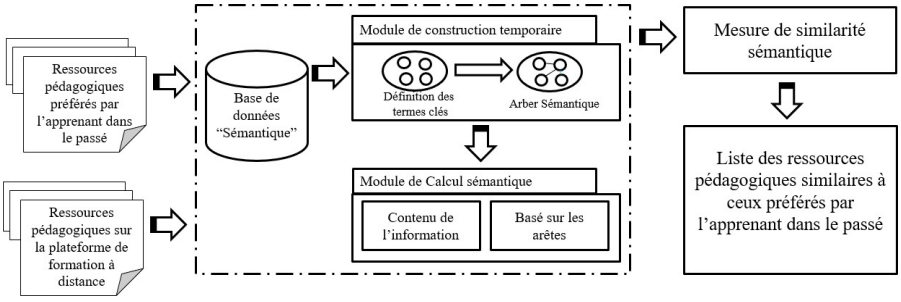

Ressources pédagogiques préférés par l'apprenant dans le passé
Ressources pédagogiques sur la plateforme de formation à distance
Base de données "Sémantique"
Module de construction temporaire
Définition des termes clés
Arber Sémantique
Module de Calcul sémantique
Contenu de l'information
Basé sur les arêtes
Mesure de similarité sémantique
Liste des ressources pédagogiques similaires à ceux préférés par l'apprenant dans le passé

Figure 3. Diagramme de calcul de similarité sémantique

3.4.1. Phase 1 : Module de construction temporaire

L'objectif du module est de sélectionner tous les mots du texte qui existent sur WordNet et d'obtenir la relation entre ces mots. Nous utilisons WordNet pour générer une représentation de texte plus riche. Dans ce module, nous avons utilisé les hyperonymes fournis par WordNet comme des fonctionnalités utiles pour l'analyse de textes.

3.4.2. Phase 2 : Module de calcul sémantique

Nous utilisons l'algorithme longueur de chemin qui utilise la mesure de similarité sémantique pour trouver les sens appropriés des mots selon le contexte à l'échelle de la phrase ou du texte.

Lorsque les concepts sont organisés dans une hiérarchie, il est convenable de mesurer la similarité en fonction de mesures structurelles qui trouvent des longueurs de chemin entre les concepts. Rada *et al.* (1989) ont développé une mesure basée sur la longueur des chemins entre les concepts dans la hiérarchie WordNet. La mesure du plus court chemin met l'accent sur la proximité de deux concepts dans la hiérarchie.

Dans un graphe de hiérarchie de thésaurus, plus le chemin entre deux mots est court, plus ces mots sont similaires :

- les mots sont assez similaires aux parents ;
- les mots sont moins similaires aux mots qui leur sont loin dans le réseau ;
- Pathlen(c1, c2) = nombre d'arêtes du plus court chemin ;
- la similarité basée sur le chemin implique souvent une transformation logarithmique ;
- la similarité basée sur la longueur du chemin est :

$$simpath(c1, c2) = -\log pathlen(c1, c2)$$

3.4.3. Phase 3 : Procédures de mesure de similarité sémantique

Les vecteurs sémantiques pour les phrases T_1 et T_2 peuvent être formés à partir des statistiques de T et du corpus, où $T = T_1 \cup T_2$, T_1={$w_{1,1}$, $w_{1,2}$,……} et T_2={$w_{2,1}$, $w_{2,2}$, ….}.

Le processus de dérivation de vecteurs sémantiques pour T_1 sera comme suit :

$$Word\ w, define\ sim(w_{1,1}, w_{2,1}) = max_{c1,c2}[sim(c1, c2)]$$

$$Sim(T_1, T_2) = \sum_{i=1}^{n} \frac{Sim(W_{1,i}, W_{2,i})}{n}$$

4. Expérimentation et résultats

Comme annoncé dans l'introduction, le système de recommandation que nous proposons est basé sur l'analyse des traces d'apprentissage. Les caractéristiques de la population visée sont : Apprenants inscrits en master d'informatique :

Nombre : 100 apprenants

Nombre de cours de master : 24 cours

Parcours : master 1 et master 2

Pour tester le système de recommandation, toutes les traces d'apprentissage concernant les tests de connaissances, les préférences et styles d'apprentissage ont été extraits. A partir de ces données, nous avons créé une base de données et développé une application (Figure 1) qui permet de gérer les recommandations de cours. Pour rappel, les recommandations sont des cours en ligne de la plateforme de MOOC.

Afin d'évaluer notre approche, nous avons adopté comme première étape une évaluation qualitative, où les 100 apprenants ont été sélectionnés et plus de 24 ressources pédagogiques. Ces ressources sont liées à différents domaines de connaissances de l'informatique. Nous avons établi un questionnaire d'évaluation avec une dizaine de questions en utilisant le score de Likert (Boone *et al.*, 2012) sur une échelle des cinq points (A : Très satisfait ; B : Plutôt satisfait ; C : Ni satisfait, ni insatisfait ; D : Plutôt insatisfait ; E : Très insatisfait) (Boone *et al.*, 2012). Les questions sont classées selon deux catégories : efficacité, en termes de pertinence des résultats de recommandation et utilité, en termes de satisfaction d'apprentissage et d'intention de réutilisation des ressources suggérées.

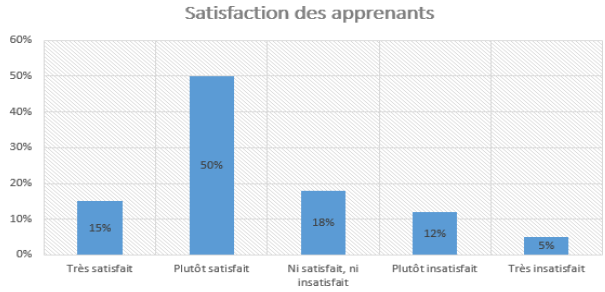


Figure 4 . Score du questionnaire sur l'échelle de Likert

Une évaluation qualitative de notre approche montre son intérêt pour les apprenants. Cependant et afin de tester notre approche plus profondément, des

expérimentations dans des situations réelles de formation sont nécessaires, en demandant à des apprenants d'intégrer cette approche dans leurs pratiques quotidiennes, afin de mieux voir les avantages et les limites du système de recommandation.

5. Conclusion

Dans cet article, nous proposons un système de recommandation personnalisé combinant les techniques de profilage des apprenants, d'évaluation des connaissances, style d'apprentissage et d'analyse des traces, afin d'améliorer le processus d'apprentissage en recommandant des ressources pédagogiques (cours, séquences de cours, vidéos, documents, etc.). Nous proposons un système de recommandation basée sur le filtrage collaboratif et le filtrage basé sur le contenu proposant automatiquement des recommandations personnalisées pour chaque apprenant. L'approche a été ensuite testée sur une situation réelle, qui a montré son efficacité et son utilité. En perspective au développement de ce projet, d'une part, nous prévoyons de recommander des groupes de travail. D'autre part, nous envisageons d'améliorer notre système de recommandation par une méthode de classification hiérarchique pour classer les ressources pédagogiques (documents, cours, section, vidéo, etc.) pour modéliser les besoins d'apprentissage.

Références

Gediminas A., Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. Transactions on Knowledge and Data Engineering. 7(6):734-749.

Baidada M., Mansouri K., Poirier F. (2019). Personalized E-learning Recommender System to Adjust Learners' Level. In EdMedia+ Innovate Learning, Association for the Advancement of Computing in Education, pp. 1353-1357.

Badru S. J., Karypis G., Konstan J., Riedl T. J. (2000). Analysis of recommendation algorithms for e-commerce. ln Proceedings of the 2nd ACM conference on Electronic commerce: ACM.

Berrut C., Nathalie D. (2003). Filtrage collaboratif. In Assistance intelligente à la recherche d'informations, Hermes Science Publications, p. 30.

Boone H. N., Boone D. A. (2012). Analyzing likert data. Journal of extension, 50(2):1-5.

Bulut B., Kaya B., Alhajj R., Kaya M. (2018). A Paper Recommendation System Based on User's Research Interests. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, IEEE, pp. 911-915.

Chekkai N., Chikhi S., Kheddouci H. (2011). Nouvelle Approche à base de Graphes pour les Systèmes de Recommandation Collaboratifs. In CIIA.

Cohen W.W. (1996). Learning rules that classify e-mail/ AAAI Spring Symposium on Machine Learning in information Access.

Coates H., James R., Baldwin G. (2005). A critical examination of the effects of learning management systems on university teaching and learning. Tertiary education and management, Vol. 11, pp. 19-36.

Downes S. (2012). Learning objects: Resources for learning worldwide. In Online education using learning objects, Routledge, pp. 47-57.

Drachsler H., Verbert K., Santos O. C., Manouselis N. (2015). Panorama of recommender systems to support learning. In Recommender systems handbook, Springer, Boston, MA, pp. 421-451.

Du Y., Ranwez S., Sutton-Charani N., Ranwez V. (2019). Apports des ontologies aux systèmes de recommandation: état de l'art et perspectives.

Felder R. M. (1996). Matters of style. ASEE prism, 6(4):18-23.

Ghézala H., Kboubi F., Gorrab A. (2019). Social Information Retrieval and Recommendation: state-of-the-art and future research. Revue Africaine de la Recherche en Informatique et Mathématiques Appliquées, Vol. 27.

Guy I., Zwerdling N., Ronen I., Carmel D., Uziel E. (2010). Social media recommendation based on people and tags. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp. 194-201.

Lahlou F. Z., Benbrahim H., Kassou I. (2017). Context Aware Recommender System Algorithms: State of the Art and Focus on Factorization Based Methods. Electronic Journal of Information Technology

Malone T., Grant K., Turbak F., Brobst S., Cohen M. (1987). Intelligent information sharing systems. Communications of the ACM, 30(5):390–402.

Mestrovic A., Calì A. (2016). An ontology-based approach to information retrieval. In Semanitic Keyword-based Search on Structured Data Sources, pp. 150-156. Springer, Cham.

Mladenic D. (1999). Text-learning and related intelligent agents: A survey. IEEE Intelligent Systems, 14(4):44-54.

Ndiyae N. M., Chaabi Y., Lekdioui K., Lishou C. (2019). Recommending system for digital educational resources based on learning analysis. In Proceedings of the New Challenges in Data Sciences: Acts of the Second Conference of the Moroccan Classification Society, ACM, pp. 15.

Negre E. (2018). Les systèmes de recommandation : une catégorisation.

O'Keefe J. (1979). A review of the hippocampal place cells. Progress in neurobiology, 13(4):419-439.

Sudha S. (2016). WordNet–An On-line Lexical Database.

Pazzani M. Billsus D. (2007). The Adaptive Web, Vol. 4321/2007 de Lecture Notes in Computer Science, chapitre Content-Based Recommendation.

Rada R., Mili H., Bicknell E., Blettner M. (1989). Development and application of a metric on semantic nets. IEEE transactions on systems, man, and cybernetics, 19(1):17-30.

Vizeshfar F., Torabizadeh C. (2018). The effect of teaching based on dominant learning style on nursing students' academic achievement. Nurse education in practice, Vol. 28, pp. 103-108.

Wan S., Niu Z. (2018). An e-learning recommendation approach based on the self-organization of learning resource. Knowledge-Based Systems, Vol. 160, pp. 71-87.

Zaier Z. (2010). Modèle multi-agents pour le filtrage collaboratif de l'information. Thèse de doctorat de l'université de Québec, Montréal, Canada.

DiLiTAL'19, p. 41-53

Paramètres optimaux pour la reconnaissance automatique des chiffres et lettres amazighes à base de Hidden Markov Model Toolkit

Safâa EL OUAHABI, Mohamed ATOUNTI, Mohamed BELLOUKI

Laboratoire MASI, Faculté Pluridisciplinaire de Nador, Université Mohamed Premier Oujda

safaa.elouahabi@gmail.com, atounti@hotmail.fr, mbellouki@hotmail.com

Résumé.

Avec la généralisation de l'utilisation des ordinateurs numériques, il devient de plus en plus nécessaire de pouvoir communiquer avec des machines de manière simple. L'une des tâches principales qui simplifie la communication avec des machines est la reconnaissance vocale. Cet article concerne la Reconnaissance Automatique (RA) des chiffres et lettres amazighes parlés, en utilisant la plate-forme Hidden Markov Model Toolkit. Un certain nombre de choix sont faits sur ce système, comme le nombre d'états des Modèles de Markov Cachés, le type de densités de probabilité d'émission associées aux états (Modèle de Mélange Gaussiens), et la représentation du signal par des coefficients spectraux et des coefficients différentiels. Les résultats de la reconnaissance de notre système le situent à un niveau de performance comparable à celui atteint par les systèmes de reconnaissance Markoviens.

1. Introduction

La parole est certainement le moyen le plus direct et le plus naturel utilisé par l'homme pour échanger l'information. Le progrès enregistré dans le domaine du traitement du signal, le développement des moyens informatiques (matériels et logiciels) et l'apport de l'intelligence artificielle permettent d'envisager l'utilisation de la parole pour communiquer et dialoguer avec une machine. Cet avantage a motivé plusieurs chercheurs à projeter leur attention sur la conception des systèmes reconnaissant la parole. Ces systèmes sont connus par le nom de Système de Reconnaissance Automatique de la Parole (SRAP).

Grâce aux progrès récents de la puissance informatique, de nombreuses techniques de reconnaissance de la parole ont été également améliorées et perfectionnées, notamment pour les langues ou les dialectes. Nous proposons dans cette contribution une approche basée sur les Modèles de Markov Cachés (MMCs) en combinaison avec les Modèle de Mélange Gaussiens (MMGs). L'utilisation de MMCs pour la reconnaissance de la parole est devenue prédominante depuis plusieurs années, car un signal de parole peut être vu comme un signal stationnaire de courte durée. Les MMCs peuvent être formés automatiquement et ils sont simples et faciles à utiliser. Les raisons pour lesquelles cette méthode est devenue si populaire sont le cadre statistique inhérent (mathématiquement précis), la facilité et la disponibilité des algorithmes d'entraînement permettant d'estimer les paramètres des modèles à partir d'ensembles d'entraînement finis de données de parole. La flexibilité du système de reconnaissance qui en résulte, où nous pouvons facilement changer la taille, le type ou l'architecture des modèles en fonction des mots, les sons, etc.

Nous décrivons, dans cet article, les différentes phases adoptées dans le processus de mise en place de notre système de reconnaissance de chiffres et de lettres amazighes à base de MMCs. La première phase est celle de la préparation du corpus et son étiquetage. La deuxième concerne la procédure d'extraction des caractéristiques. Dans la troisième phase, nous avons utilisé un classifieur analytique de type Markovien où chaque ensemble a des vecteurs de caractéristiques, suivant une distribution de densité continue par la loi gaussienne. Et à la fin du processus, le test et l'analyse des résultats expérimentaux ont été réalisés.

2. Travaux antérieurs sur la reconnaissance des chiffres et lettres amazighes

En général, les chercheurs dans le domaine de la RAP ont ciblé des systèmes à petit, moyen et grand vocabulaire, continue ou discrète pour différentes langues (Kimutai *et al.*, 2013 ; Kumar *et al.*, 2012 ; Sameti *et al*., 2011 ; Gales *et al*., 2007). Cependant au Maroc, beaucoup d'efforts ont été effectués pour développer un système de RAP pour la langue amazighe. Certains travaux sont menés dans cette direction pour les chiffres et les lettres amazighes. Dans cette section, nous citons les travaux publiés qui touchent la reconnaissance de la parole amazighe.

- Sadiqui *et al.* (2018) ont utilisé un modèle acoustique de la langue arabe pour en créer un pour la langue amazighe. Le système réalisé a atteint un taux de reconnaissance d'environ **73**% par mot. Le potentiel et l'efficacité de l'approche proposée sont démontrés par des expériences et des comparaisons avec d'autres approches.

▪ Telmem *et al.* (2018) ont développé un système de reconnaissance automatique de la parole pour la langue amazighe. Ce système est basé sur les modèles de Markov cachés avec des mélanges gaussiens pour générer des modèles acoustiques. Ils ont utilisé l'outil CMU Sphinx-4. La taille de la base de voix pour ce travail de recherche est de 11220 mots. Le système a atteint un taux de reconnaissance de **90**%.

▪ Dans la contribution d'Abenaou *et al.* (2014), les auteurs proposent un système de reconnaissance automatique des mots isolés de la langue amazighe basée sur les transformations orthogonales paramétrables. Suite à l'étude comparative réalisée sur le système à base des transformations orthogonales paramétrables et sur les transformations orthogonales de Walsh, Haar et Fourier, ils ont pu atteindre un taux de reconnaissance plus élevé qui atteint **96**%.

▪ El Ghazi *et al.* (2014) ont exploité les règles de construction des chiffres amazighes pour construire un système de reconnaissance automatique de la parole. Ce système est basé sur la synthèse de chiffres enchaînés à partir d'un nombre isolé de 1 à 10. Ce système permet également d'étendre la reconnaissance automatique à l'amazighe. Les résultats obtenus sont satisfaisants par rapport à la taille de la base de données d'apprentissage. Par conséquent, leur système constitue une première partie d'un système de sécurité.

▪ Satori *et al.* (2014) ont étudié l'indépendance du système de reconnaissance de la parole en utilisant une base de voix « alphadigits » correspondant aux chiffres et lettres prononcés par des locuteurs d'origine amazighe. Ce système a été mis en œuvre, en utilisant CMU Sphinx qui est basé sur les MMCs. Les résultats de reconnaissance montrent que leur système est indépendant et sa performance est satisfaisante.

▪ El Ghazi *et al.* (2011) ont présenté un système de reconnaissance automatique de la parole amazighe basé sur la transcription en alphabet tifinaghe reconnu par l'Institut Royal de la Culture Amazighe (IRCAM). Ils ont utilisé le MMC et une comparaison avec la méthode de programmation dynamique. Les résultats obtenus par le modèle de Markov caché sont très satisfaisants, malgré le nombre limité de locuteurs et de la taille de la base de données. Ceci montre l'importance de modélisation stochastique et probabiliste dans le domaine de la reconnaissance.

3. Fondements théoriques

3.1. Modèles de Markov Cachés

Les MMCs sont des outils statistiques permettant de modéliser des phénomènes stochastiques. La modélisation d'un modèle de Markov caché est largement détaillée et illustrée mathématiquement par Rabiner (1989). Leur principe de modélisation est de mesurer une variable aléatoire $X(t)$ et de chercher à modéliser le processus par un modèle stochastique (MMC) défini par $\boldsymbol{M(A,B,\Pi)}$, comme illustré sur la Figure 1.

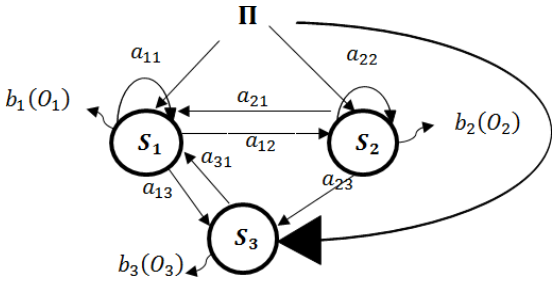


Figure 1. Représentation d'une chaîne de Markov Caché

Sachant que :

- $\boldsymbol{A} = \{a_{ij} = \mathbb{P}(q_t = S_j | q_{t-1} = S_i)\}$ est la matrice des probabilités de transitions entre les états, avec $1 \leq i,j \leq N$, où N indique le nombre d'états de la chaîne de Markov S={S_1, S_2, ….., S_N} et q_t indique l'état à l'instant *t*.
- $\boldsymbol{B} = \{b_i(O_k) = \mathbb{P}(O_k | q_t = S_i)\}$ est la matrice des probabilités d'émission des observations dans chaque état, avec $1 \leq i \leq N$, $1 \leq k \leq M$ et $\vartheta = \{O_1, O_2, \ldots, O_M\}$ sont des symboles émis par les états .
- $\boldsymbol{\Pi} = \{\pi_i\} = \{\mathbb{P}(q_1 = S_i)\}$ est le vecteur des probabilités initiales des états, où chaque état S_i la probabilité d'être atteint à partir de l'état initial q1 est π_i et $1 \leq i \leq N$.

3.2. MMG-MMC pour la modélisation et la reconnaissance de la parole

En reconnaissance de la parole, l'une des approches d'apprentissage génératif la plus répandue repose sur les MMCs en combinaison avec le Modèle MMG (Dempster *et al.*, 1977 ; Deng *et al.*, 1991a ; Juang *et al.*, 1986 ; Rabiner, 1989 ; Rabiner *et al.*, 1993). Un MMG-MMC est un modèle statistique qui décrit deux processus aléatoires dépendants, un processus observable et un processus caché, processus de Markov. La séquence d'observation est supposée être générée par chaque état caché selon une distribution de mélange gaussien. Un MMG- MMC est paramétré par un vecteur de probabilités d'état, la matrice de probabilité de transition d'état et par un ensemble de paramètres dépendant de l'état dans les modèles de mélange gaussien. En termes de modélisation de la parole, un état dans le MMC-MMG est généralement associé à un sous-segment d'un phonème. Une innovation importante dans l'utilisation des MMCs pour la reconnaissance de la parole est l'introduction d'états dépendants du contexte (Deng *et al.*, 1991b ; Huang *et al.*, 2001). Cette dépendance de contexte joue un rôle essentiel dans la reconnaissance de la parole dans la partie de l'apprentissage (Chen *et al.*, 2012 ; Dahl *et al.*, 2011; Dahl *et al.*, 2012 ; Seide *et al.*, 2011 ; Yu *et al.*, 2010).

4. Système de reconnaissance de la parole amazighe à base de Hidden Markov Model Toolkit

Dans cette section, nous présentons l'architecture générale du système proposé avec une description détaillée des différentes techniques utilisées, y compris la procédure d'extractions des caractéristiques, la modélisation analytique de type Markovienne. Enfin, nous présentons les résultats obtenus par les différents ensembles MMCs et MMG.

Pour élaborer notre système, nous avons utilisé la boite à outils HTK (Hidden Markov Model Toolkit), qui est dédiée aux MMCs pour l'apprentissage et la reconnaissance des modèles. Sa réalisation consiste d'abord en la constriction des modèles phonétiques indépendants du contexte. Un certain nombre de choix sont faits pour ce système, comme le nombre d'états des modèles, le type de densité de probabilité d'émission associée aux états et la représentation du signal par des coefficients spectraux et des coefficients différentiels.

Les principales composantes de notre système sont illustrées à la Figure 2 :

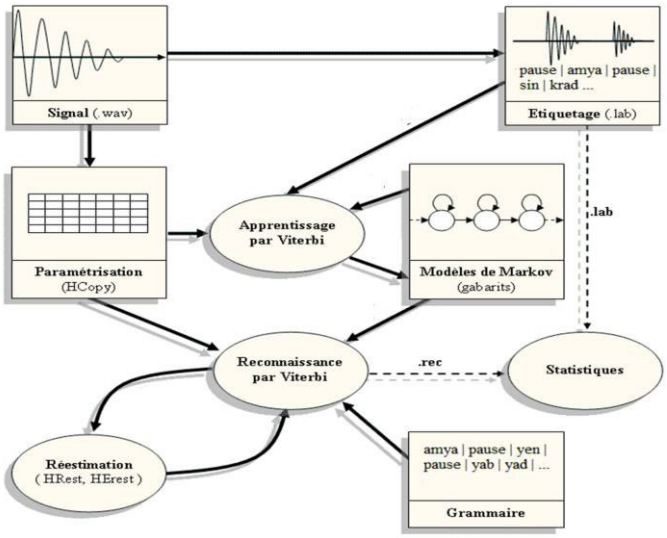


Figure 2. Architecture du système de reconnaissance basé sur HTK

4.1. Corpus du signal de la parole

Le corpus comprend 10 chiffres et 33 lettres amazighes. Chaque chiffre et lettre est prononcé 10 fois dans des fichiers de données séparés, où chaque fichier contient 1 énoncé. La base se compose de 19 500 enregistrements de format .wav (Voir Tableau 1).

Data	Données d'apprentissage	Données de Test	Total
Chiffres	2000	1000	3000
Lettres	13200	3300	16500
Total	15200	4400	**19500**

Tableau 1. Corpus de chiffres et de lettres amazighes

4.2.Etiquetage

Au cours des sessions d'enregistrement de la parole, chaque locuteur prononce la liste de mots. Chaque suite de mots a une transcription phonétique qui ne fournissait qu'une séquence phonétique correspondant à la séquence des mots, c'est-à-dire qu'il n'y avait pas d'indice de temps. Par conséquent, cette phase nécessite beaucoup de temps pour terminer une transcription. Chaque fichier étiqueté est enregistré au format.lab qui est un fichier texte simple. Ces fichiers sont utilisés dans la phase de génération du modèle acoustique du système. Wavesurfer permet d'afficher un fichier audio sur une interface, puis étiqueter les zones importantes en sélectionnant leurs parties associées (Figure 3). Chaque fichier wav dans le corpus est étiqueté avec wavesurfer, ce qui entraîne 19 500 fichiers (.lab).

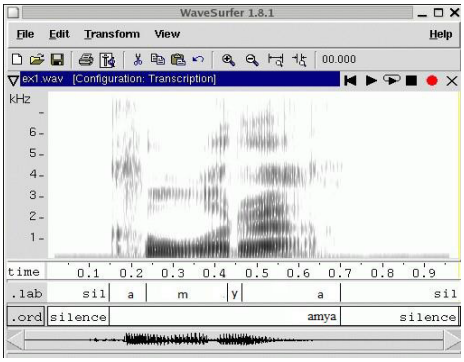


Figure 3. Etiquetage du signal du mot amya

4.3. Paramétrisation

L'extraction des caractéristiques est responsable de la transformation du signal de la parole en un flux de vecteurs de caractéristiques qui n'ont que l'information requise pour identifier un énoncé donné. La technique d'analyse spectrale la plus largement utilisée pour l'extraction de vecteurs caractéristiques est Mel-Frequency Cepstrum Coefficient (MFCC), qui est utilisée pour imiter l'oreille humaine.

parametrisation.conf

```
SOURCEFORMAT = WAVE
#TARGETRATE = 100000
#WINDOWSIZE = 250000      # 25 ms
#USEHAMMING = T           # la FFT utilisera un fenêtrage de Hamming
#TARGETKIND = MFCC_0_D_A   # 3 MFCC + 3 delta MFCC + 2 Energie
#NUMCHANS = 8             # taille d'un vecteur
#ENOMALISE = T            # normalisation de l'energie sur le fichier audio
#CEPLIFTER = 22
#NUMCEPS = 3
#PREEMCOEF = 0.97         # coefficient de pre-emphase
#SAVECOMPRESSED = F
#SAVEWITHCRC = F
NONUMSCAPES = T
TARGETKIND = MFCC_0_D_A
TARGETRATE = 100000.0
SAVECOMPRESSED = T
SAVEWITHCRC = T
WINDOWSIZE = 250000.0
USEHAMMING = T
PREEMCOEF = 0.97
NUMCHANS = 26
CEPLIFTER = 22
NUMCEPS = 12
ENORMALISE = F
```

Figure 4. Paramètres de configuration Hcopy

Les paramètres nécessaires pour l'analyse acoustique sont : le format des fichiers vocaux d'entrée, caractéristiques à extraire, taille de la fenêtre, fonction de la fenêtre, nombre de coefficients spectraux, coefficients de préaccentuation, nombre de filtres canaux de banc et la longueur de filtrage spectral sont fourni à la commande HCopy de HTK dans un fichier de configuration parametrisation.conf. Les valeurs de paramètres utilisés dans nos différentes expériences sont montrées dans la Figure 5.

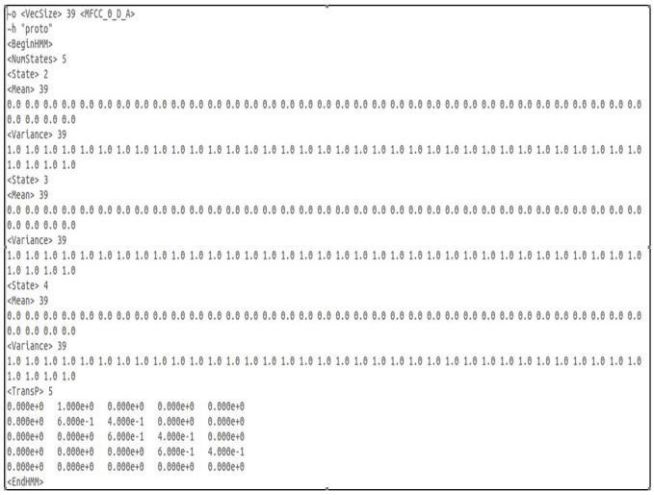


Figure 5. Définition d'un MMC à 5 états dans HTK

4.4. Estimation des paramètres et génération du modèle Acoustique

▪ Initialement, un prototype MMC doit être défini. Le but de la définition du prototype est seulement de spécifier les caractéristiques et la topologie du MMC. Les valeurs moyennes ou de variances ne sont pas importantes dans un prototype. Les paramètres réels seront calculés par les outils d'entraînement de manière incrémentale.

▪ La topologie principale proposée dans HTK est une topologie MMC à 5 états (voir Figure 5). Les états non émis n'apparaissent pas en tant que définition d'état dans un fichier de définition MMC. Leur effet ne peut être vu que dans la matrice de probabilité de transition. La dernière ligne et la première colonne de cette matrice sont mises à '0'.

▪ L'estimation des paramètres de ces modèles progresse de façon incrémentale. Au début du processus de l'apprentissage, le modèle de silence est le même que les autres modèles de phonème. Progressivement, le modèle de silence est modifié et le modèle de pause courte est inséré dans le processus d'apprentissage dans les étapes ultérieures.

▪ Apprentissage : Chaque modèle doit être appris, ainsi les moyennes, les variances et les probabilités de transition entre états sont réestimées jusqu'à ce qu'un seuil de convergence ou qu'un nombre maximum d'itérations soit atteint. Ceci est fait par l'algorithme de Viterbi.

4.5. Modèle de langage et décodage

▪ La reconnaissance se fait avec l'algorithme de Viterbi qui, dans l'environnement HTK, nécessite l'utilisation de diverses connaissances : les modèles HMM, le modèle de langage et le dictionnaire.

▪ Le modèle de langage est généré par la fonction HParse à partir de la description du réseau dans un fichier texte.

4.6. Expériences et résultats

Nous avons essayé d'optimiser les différents paramètres déterminants dans la performance des systèmes tels que le nombre d'états du MMC, les distributions de probabilités, la quantité de données d'apprentissage. Nous avons, fait varier le nombre d'états pour chaque modèle de mot dans un intervalle [3,6] avec un pas d'un seul état par modèle. Nous avons utilisé pour ces expériences, la base décrite dans la section 4.1.

La première expérience consiste à reconnaître les 10 chiffres amazighes. La Figure 6 montre le taux de reconnaissance en utilisant des combinaisons d'états MMCs {3, 4, 5, 6} et divers nombres de distributions gaussiennes {2, 4, 8, 16, 32}.

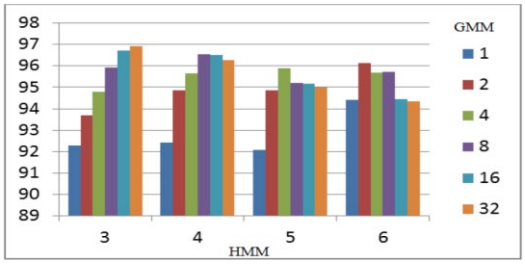


Figure 6. Taux de reconnaissance (%) pour différents MMCs et MMGs pour les chiffres amazighes

Les résultats expérimentaux montrent que généralement le système a donné de bonnes performances. En considérant le meilleur taux de reconnaissance, les meilleurs paramètres choisis sont les suivants : MMC = 3 états avec la combinaison de 32 MMG pour les chiffres. La meilleure performance du système est de 96, 93%.

La seconde expérience consiste à reconnaître 33 lettres amazighes. La Figure 7 montre le taux de reconnaissance en utilisant des combinaisons d'états MMCs {3, 4, 5, 6} et divers nombres de distributions gaussiennes {2, 4, 8, 16, 32}.

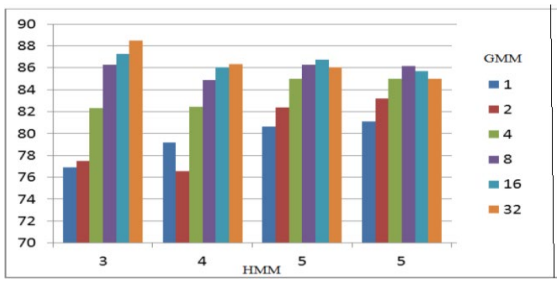


Figure 7. Taux de reconnaissance (%) pour différents MMCs et MMGs pour les lettres amazighes

Les résultats montrent que le système offre une bonne performance de 88,48% en utilisant MMC = 3 états avec une combinaison de 32 mélanges gaussiens pour la base de données des lettres.

Dans le cas de la troisième expérience, les deux bases de données ont été combinées. Les résultats des tests sont détaillés dans le Tableau 2.

MMC = 3	MMGs				
	2	4	8	16	32
Taux de reconnaissance total (%)	80,83	85,01	89	89,82	**91,31**

Tableau 2. Taux de reconnaissance (%) pour différents MMCs et MMGs pour les lettres et les chiffres amazighes

Les performances du système sont de 85,01%, 89,89%, 82% et 91,31% pour l'utilisation de distributions de mélanges 4, 8, 16 et 32 gaussiennes respectivement. Nous constatons que 32 mélanges de gaussiennes a permis d'obtenir le meilleur taux de reconnaissance d'ordre 91,31%.

4.7. Comparaisons des systèmes

Le résultat obtenu nous permet de comparer nos résultats avec ceux des autres systèmes portant sur la reconnaissance automatique de la parole amazighe. Cette comparaison est basée sur les conditions suivantes : la modélisation Markovienne, la distribution gaussienne, la taille de la base de données. Avec différence dans le type du décodeur utilisé Sphinx Vs HTK. Le Tableau 3 montre l'analyse comparative de la recherche faite dans le domaine de la reconnaissance de chiffres et lettres isolés en utilisant diverses techniques.

Systèmes	**Approches et Techniques**	**Taux de reconnaissances**
Satori *et al.*, 2014	MMC+MMG =>CMUSphinx	89,07 %
El Ghazi *et al.*, 2014	DTW Vs MMC	90%
Telmem *et al.*, 2018	MMC+MMG =>CMUSphinx	90%
Notre Système	HMM + MMG => HTK	**91,31%**

Tableau 3. Comparaison du système proposé avec certains systèmes

De ce tableau et l'analyse comparative de cette section, c'est observé que les chercheurs utilisant principalement MFCC pour l'extraction des caractéristiques et MMC + MMG comme classifieur atteignant un très bon taux de reconnaissance. Cela nous permet de démontrer : l'efficacité de mélange des caractéristiques géométriques et statistiques essentiellement la modélisation par les MMCs que nous avons utilisé dans nos travaux de recherche.

5. Conclusion

Les résultats expérimentaux montrent l'efficacité et la robustesse de notre approche utilisée à base de combinaison des MMCs et MMGs. De nombreuses perspectives sont ouvertes sur différents aspects de ce travail, aussi bien au niveau de l'étape d'apprentissage que de l'étape d'extraction des caractéristiques. Au sujet de la base de données des chiffres et des lettres amazighes, nous comptons rendre cette ressource disponible pour des recherches et des utilisations académiques, vu que cette tâche de préparation de ressource est coûteuse en termes de temps et de compétences humaines. D'autre part, nous envisageons des travaux de recherches apportant des améliorations d'un côté sur les approches et d'un autre côté sur le développement des différentes applications facilitant l'intégration de la langue amazighe dans le domaine des TICs.

Références

Abenaou A., Ataa Allah F., Nsiri B. (2014). Vers un système de reconnaissance automatique de la parole amazighe basé sur les transformations orthogonales paramétrables. Asinag n° 9, pp. 133-145.

Amri S., Zenkouar L. (2019). Amazigh POS Tagging Using TreeTagger: A Language Independant Model. Advanced Intelligent Systems for Computing Sciences, Vol. 5.

Chen X., Eversole A., Li G., Yu D., Seide F. (2012), Pipelined back-propagation for context dependent deep neural networks. Proceedings of Annual Conference of International Speech Communication Association.

Dahl G., Yu D., Deng L., Acero A. (2011). Large vocabulary continuous speech recognition with context-dependent DBN-HMMs. Proceedings of International Conference on Acoustics, Speech and Signal Processing.

Dahl G., Yu D., Deng L., Acero A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio, Speech Lang. Process, 20(1):30-42.

Dempster A. P., Laird N. M., Rubin D. B. (1977). Maximum-likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. Ser. B. 39.

Deng L., Lennig M., Seitz F., Mermelstein P. (1991a). Large vocabulary word recognition using context-dependent allophonic hidden Markov models. Comput. Speech Lang 4, pp. 345-357.

Deng L., Kenny P., Lennig M., Gupta V., Seitz F., Mermelsten P. (1991b). Phonemic hidden Markov models with continuous mixture output densities for large vocabulary word recognition. IEEE Trans. Acoust. Speech Signal Process. 39(7):1677-1681.

El Ghazi A., Daoui C., Idrissi N. (2012). Automatic Speech Recognition system concerning the Moroccan dialect (Darija and Tamazight). International Journal of Engineering Science and Technology (IJEST), Vol. 4, n°3.

El Ghazi A., Daoui C., Idrissi N. (2014). Automatic Speech Recognition for Tamazight Enchained Digits. World Journal Control Science and Engineering, 2(1):1-5.

Gales M. J., Diehl F., Raut C. K., Tomalin M., Woodland P. C., Yu K. (2007). Development of a phonetic system for large vocabulary Arabic speech recognition. Proceedings of Automatic Speech Recognition & Understanding, (ASRU), IEEE Workshop, 9-13 Décembre 2007, Kyuto, Japan, pp. 24-29.

Huang X., Acero A., Hon H. W., Reddy R. (2001). Spoken Language Processing, Vol. 18. Prentice Hall, Englewood Cliffs.

Juang B. H., Levinson S. E., Sondhi M. M. (1986). Maximum likelihood estimation for mixture multivariate stochastic observations of Markov chains. IEEE Int. Symp. Inf. Theory, 32(2): 307-309.

Kimutai S. K., Milgo E., Gichoya D. (2013). Isolated Swahili Words Recognition using Sphinx4. International Journal of Emerging Science and Engineering (IJESE), Vol.2, Issue 2.

Kumar K., Jain A., Aggarwal R. K. (2012). A Hindi speech recognition system for connected words using HTK. International Journal on Computational Systems Engineering, 1(1):25-32.

Rabiner L. R. (1989). A tutorial on hidden Markov models and select applications, Speech Recognition Revue, 77(2):257-286.

Rabiner L. R., Juang B. H. (1993). Fundamentals of Speech Recognition. Prentice- Hall, Englewood Cliff, New Jersy.

Sadiqui A., Zinedine A. (2018). Towards a Speech Recognizer for Multiple Languages Using Arabic Acoustic Model Application to Amazigh Language. International Conference on Arabic Language Processing, pp. 67-78.

Sameti H., Veisi H., Bahrani M., Babaali B., Hosseinzadeh K. (2011). A large vocabulary continuous speech recognition system for Persian language, EURASIP Journal on Audio, Speech, and Music Processing.

Satori H., El Haoussi F. (2014). Investigation Amazigh speech recognition using CMU tools. International Journal of Speech Technology, 17(3): 235-243.

Seide F., Li G., Chen X., Yu D. (2011). Feature engineering in context-dependent deep neural networks for conversational speech transcription. Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 24–29.

Telmem M., Ghanou Y. (2018). Estimation of the Optimal HMM Parameters for Amazigh Speech Recognition System Using CMU-Sphinx. Procedia Computer Science, Vol. 127, pp. 92-101.

Yu D., Deng L., Dahl G. (2010). Roles of pre-training and fine-tuning in context-dependent DBNHMMs for real-world speech recognition. NIPS Workshop on Deep Learning and Unsupervised Feature Learning.

DiLiTAL'19, p. 55-65

Free large-scale Arabic speech corpus for speech synthesis

Oumaima ZINE, Abdelouafi MEZIANE

Computer science department, Mohammed First University, Oujda, Morocco

zine.oumaima@gmail.com,
abdelouafi_meziane@yahoo.fr

Abstract.

One of the key issues in building a high-quality text-to-speech system is the speech corpus design. The robustness of the recent speech synthesis approaches relies on large-scale and prosodically balanced speech databases. As such corpora are unavailable for the Arabic language, we have developed one for this purpose. It aimed at building unit-selection based Text-to-Speech systems for Modern Standard Arabic, as for the most advanced state of the art speech synthesis systems, namely neural end-to-end speech synthesis. Thus, we discuss the different stages undertaken for the design and construction process. The corpus consists of 81 thousand words fully diactritized that required 15 hours of speech data. In this paper, the attention is given mainly to (1) text collection, taking into consideration both phonetically and prosodically balanced sentences, (2) text pre-processing, (3) recording stage, in addition to the orthographic and the phonetic level annotation.

1. Introduction

According to (Choukri and Hamid, 2005), the required speech data volume for a concatenation-based Text-to-Speech (TTS) is 5 hours. However, this assumption is restricted to short speech segments, typically diphones or triphones, with only one instance of each type.

In our previous work, we introduced a speech synthesis approach based on the concatenation of a novel speech unit called lemma-like. We proved that this latter helped to enhance the naturalness of the produced speech. Given a large number of Arabic lemmas, the baseline speech segments were limited to the top 1,000 frequent lemmas of the Modern Standard Arabic (MSA). In fact, this number was found to provide a coverage of 79% of the Arabic

words, which required 4 hours of speech from which we extracted the speech database. Nonetheless, expanding the lemma-like lexicon may lead a priori

to higher coverage of the Arabic words. Hence, a great enhancement of the produced speech naturalness will be achieved. Thus, one of the main perspectives mentioned in our previous work was to target the top 5,000 frequent lemmas of MSA. However, this number of lemmas would require a large-scale Arabic speech corpus to achieve good coverage on the related speech segments, in various prosodic contexts.

Furthermore, it is assumed that lemmas are mainly dependent on domains. This implies that the coverage of the most frequent lemmas in the Arabic language would require corpora that include a broad range of domains, which are unfortunately unavailable for this purpose.

On the other hand, the linguistic structures of the written language (e.g. newspapers) differ from those found in speech, as written language tends to be more intricate than speech. In addition to some grammatical constructions that are only used in writing, as are some kinds of vocabulary.

For this purpose, the baseline corpus should represent both types of data: text derived from written scripts and text derived from transcribed speech.

Including all the considerations mentioned above, we have collected and designed a speech corpus that will fully meet the requirements of our TTS approach, as well as the state of art speech synthesis approaches. We expect that a large-scale free Arabic speech corpus that is rich in terms of vocabulary and prosodic variations would accelerate related research such as on corpus-based and end-to-end speech synthesis.

In this paper, we describe the process of collecting data and recording 15 hours of speech read by a native Arabic speaker. The corpus is designed to include different-domain utterances, such as newspapers, television broadcasts, novels, religious texts, etc.; in addition to all the different phonetic and prosodic variations - i.e. short and long sentences, dialogues, and questions. Three stages are emphasized in the following sections: the collection of the textual database, its recording and the transcription of the acoustic database thus obtained.

2. Related work

The quality of speech synthesis systems crucially depends on the richness of the speech unit database in terms of the phonetic and prosodic variations.

An overlook on the existing Arabic speech resources shows the lack of available single speaker speech corpora for MSA. However, there are free large-scale corpora for non-Arabic languages such as:

LJ Speech[1]: a public domain English speech dataset consisting of 13,100 short audio sample of a single speaker reading texts from 7 non-fiction books. A transcription is provided for each audio sample. The dataset has a total length of approximately 24 hours comprised of non-professional audiobooks from the LibriVox project and distributed under a non-restrictive license.

LibriTTS[2]: the largest English TTS corpus, designed by (Zen *et al.*, 2019) to train neural multispeaker end-to-end TTS models. The corpus consists of a set of audiobooks derived from the original materials of the LibriSpeech corpus (Panayotov *et al.*, 2015). The dataset has about 585 hours of speech data from 2,456 (1,185 female / 1,271 male) speakers and the corresponding texts. The corpus is distributed under a non-restrictive license.

SIWIS[3]: consists of 10 hours of speech. Its primary target is speech synthesis. It also contains sentences with an emphasis on specific words, in many contexts. The data is available with no restrictions on the usage, under Creative Commons Attribution 4.0 International license.

By contrast, most of the available Arabic speech datasets are directed towards Automatic Speech Recognition (ASR) purposes, whereas TTS corpora are mostly either limited to a specific domain, multi-speaker, expenses paid, and or exclusively reserved for subscribers.

As the best of our knowledge, few efforts have been made to build TTS speech corpora for the Arabic language such as:

- A free MSA speech corpus was designed and constructed in (Halabi, 2016). The corpus provides single speaker recordings and intended for parametric and unit selection speech synthesis. The transcripts were scraped from the Aljazeera learn website (Al Jazeera, 2015), which contains fully diacritized text. Afterward, a set of 884 sentences were nominated using a modified greedy selection algorithm that applies certain criteria to reduce the initial orthographic transcript size and achieve the optimal selection, taking

[1] Available at https://keithito.com/LJ-Speech-Dataset/
[2] Available at http://www.openslr.org/60/
[3] Available at http://datashare.is.ed.ac.uk/download/DS_10283_2353.zip

into consideration the full coverage of diphone units. The corpus segmentation was achieved using HTK flat-start forced alignment. The output alignments were then checked by linguistic experts for manual inspection and correction. The final speech database is 3.7 hours long, representing 1,780 utterances including nonsense auto-generated sentences in addition to the selected set. Although this dataset doesn't provide a large amount of data, especially for neural end-to-end TTS, it has been used to train a Tacotron-based text-to-speech (TTS) (Sharief, 2017; Wang *et al.*, 2017), thanks to its non-restrictive license. However, as this dataset's content is mainly news-related and with no prosodic variations, the synthesized speech is often monotonous, as shown in the output audio samples[4] synthesized with Tacotron.

- On the other hand, the Linguistics Data Consortium (LCD) provides a collection of MSA and dialectical Arabic speech and text datasets for research and development purposes. Amongst them an Arabic speech corpus for TTS applications that was carried out by RDI and ENSIAS under the NEMLAR project (Yaseen *et al.*, 2006). The data consists of 10 hours of reading speech, 5 hours from a female speaker and 5 hours from a male speaker. The recordings are provided with their corresponding transcriptions tagged with the Part of Speech (PoS) and annotated with specific markers (e.g. noise, unintelligible words, etc.). The prosodic markers were also added to the transcripts (e.g. pitch). The corpus was designed to target different domains, therefore, three different sub-corpora were included: (1) transcribed speech, (2) written text, (3) constructed phrases (frequent phrases, missing and rare diphones). However, this corpus is licensed and can only be used for a fee.

- Almosallam *et al*., (2013) created a Standard Arabic Single Speaker Corpus (SASSC) for MSA. It consists of 51 thousand words that required 7 hours of speech data. The dataset is divided into 9 sub-categories, namely numbers, dates and times, common phrases in Interactive Voice Response systems (IVR), and so on. All of the created subcategories were recorded in a monotonous style. Only some portions of the miscellaneous subcategory were recorded using different expressions such as sadness, joy, surprise, and questions for 15 minutes each. The corpus was segmented and aligned automatically on a phoneme level by an HMM aligner and then checked manually by computational linguists. The available version of the corpus includes the wav files, their corresponding EGG signal in 96 kHz and 16-bit resolution, the diacritized text and their transcription along with the phoneme

[4] Available at https://youssefsharief.github.io/arabic-tacotron-tts/

boundary segmentation.

3. The corpus design pipeline

3.1. Text selection features

One of the primary issues in collecting speech corpora is the availability of fully diacritized texts, as most modern digital Arabic written text is undiacritised. Thus, the size of most of the available speech datasets is often limited to a few hours of speech.

In this work, the selection of text data was not restricted to diacritized texts, since a semi-automatic diacritization phase was foreseen. Moreover, it's widely recommended that diversity of sources in speech corpora should be wide enough to cover different domains and styles (interrogative, dialogues, etc.). Furthermore, as the main purpose of designing this corpus was to target full coverage of the top 5,000 frequent lemma-like speech segments, adopted in our TTS system, the corpus was designed to contain a broad range of different domains, balanced in terms of vocabulary and prosodic variations.

Therefore, texts were chosen for inclusion according to two selection features: domain and timespan.

3.1.1. Selected domains

The text was crawled from different websites covering Classical Arabic (CA) and MSA, while maintaining the balance between the main domains - i.e. novels, news and finance, technology, science, and classical texts.

Further categories were added to the corpus to include the uncovered utterances, such as:

- Questions and dialogues: cover different prosodic variations of the Arabic interrogative clauses, including all types of questions, namely Yes/No, alternative, informational questions, etc.
- Academic: a specific vocabulary is often used in academic documents, as in Arabic grammar books.
- Autobiography and literature.

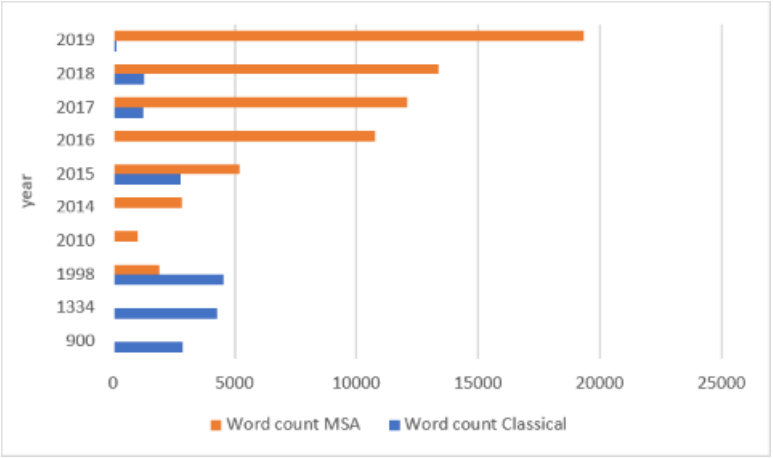


Figure1. Timespan of the collected texts according to their types (MSA or Classical)

3.1.2. Timespan

The timespan of the entire corpus varies between the year 946 - for classical scripts such as Al hadith, old books and Quranic verses - and 2019.

Figure 1 shows the timespan of the gathered texts according to their type.

3.2. Text preprocessing

The preprocessing step consists of splitting the texts into paragraphs and sentences and perform text normalization.

- Texts were first split into sentences to make it easier for the talent during recording sessions.
- Non-standard words (e.g., abbreviations, numbers and currency) words were first detected and normalized.
- The uncommon or long foreign words (e.g. names of institutions, proper names, etc.) were replaced with the tag <foreign>, in order to facilitate the recording. This is because the talent said it is difficult to remember and pronounce those words, which may result in many interruptions during the recording sessions.

4. Recording procedure

4.1. Initial speaker selection

Before selecting the speaker, we defined a set of criteria that should be taken into account during the selection process:

- A professional speaker with a pleasant voice, and fluent Arabic.

- The speaker should be available to record future extra sentences in case of speech corpus expansion or re-recording some incorrect sentences.
- The time needed to record the corpus should be kept as short as possible to minimize the effects of voice variations in time.
- The speaker's voice should not be influenced by a dominant accent.

4.2. Post speaker selection

Based on the above-mentioned criteria, we selected two candidates that satisfied the requirements. Both were given a short text to record in about 10 minutes. Afterward, the resulting speech samples of this step were analyzed. As a result, the speaker with the maximum overall rate was selected according to the following specs:

- Pronounced phonemes' boundaries are easily distinguishable in the spectrogram.
- Very low reading error rate.
- Intelligibility of the TTS resulting speech based on the concatenation of the speech segments extracted from the speaker's recordings.

4.3. Recording material

The hired male speaker is an experimented writer and an Arabic professional voice talent, with many years of experience in the field of voice-over. The speaker made the recordings in a soundproof studio, and the speech data was sampled at 24 kHz, which is according to Zen *et al.* (2019) the minimum required sampling rate for TTS. Thus, using a lower value would reduce remarkably the quality and naturalness of the synthesized speech.

In addition, the following criteria were taken into account during the recording phase:

- 16-bit depth
- utterance-level signal-to-noise ratio (SNR) > 40 dBA
- RT60 < 0.3 s
- Bandwidth: at least 40 – 20'000 Hz

4.4. Post-processing

As stated in section 3, the crawled text is partially diacritized. Thus, a full diacritization step is required before the generation of the phonetic transcription. For MSA, this process can be achieved by using automatic diacritizers. However, as the output is often not 100% accurate, it should be post-corrected manually, which can be a tedious task given the large amount of data. Apart from this, as the speaker may not follow the provided

diacritics during the recording session, the transcripts should be revised in order to align them with the speaker's pronunciation.

Therefore, we trained a model to recognize vowels in the recorded speech, using an already diacritized and checked sample. The output of the decoding step was then used to autocorrect the mistaken vowels produced by the automatic dicaritizer.

Moreover, the commas were manually annotated based on the position of silence in the recorded speech. Table 1 represents the audio duration of the recorded sub-corpora.

Subset	*Hours*
Novels	2.44
News and finance	2.43
Technology	3.04
science	2.58
Classic	4.20
Academic	0.35
literature	0.20
Questions	0.10
Total	*15.34*

Table 1. Audio duration of the different domains

5. Corpus statistics

Table 2 provides a summary of all the designed subsets. It can be seen from the number of the covered words in the six main categories- i.e. news novels, technology science and classical texts - that a certain balance between the main sub-corpora is more or less maintained. Whereas, the size of the additional categories, namely 'Questions', 'academic', and 'literature' is significantly lower than that of the main sub-corpora. This is due to the fact that the corresponding texts were reduced in order to keep only the sentences that includes uncovered lemmas, styles and prosody.

Looking at the chart represented in Figure 2, it appears that the most frequent lemmas are distributed over the different categories-i.e. novels, science, technology and news. This is due to the fact that the lemma frequency is highly correlated with the domain. That is why the construction of the lemma frequency dictionary was not restricted to a specific domain or

general topics texts. Besides, a full coverage of the common vocabulary of the different MSA domains and CA was taken into consideration.

Domain	Number of words	Distinct words	% of the total size	Avg. sentence length(words)
Novels	14.134	6923	17.54%	16.4
News and finance Sub_C_transcribed_speech Sub_C_written	12.984	6928	16.12%	23.7
Technology	14.965	7589	18.58%	22
Science Sub_C_Medecine Sub_C_Astronomy Sub_C_Earth Sub_C_Physics	14.520	6267	18.02%	19.3
Classic Sub_C_Hadith Sub_C_Verses Sub_C_Old_books	17.073	8.223	21.20%	22.9
Culture	3.966	2.310	5%	36.4
Academic	2.069	1279	2.55%	5.5
Questions	823	538	1.01%	23.9
Total	**81K**	**3.1764**	**100%**	**19.7**

Table 2. List of the corpus categories and sub-categories

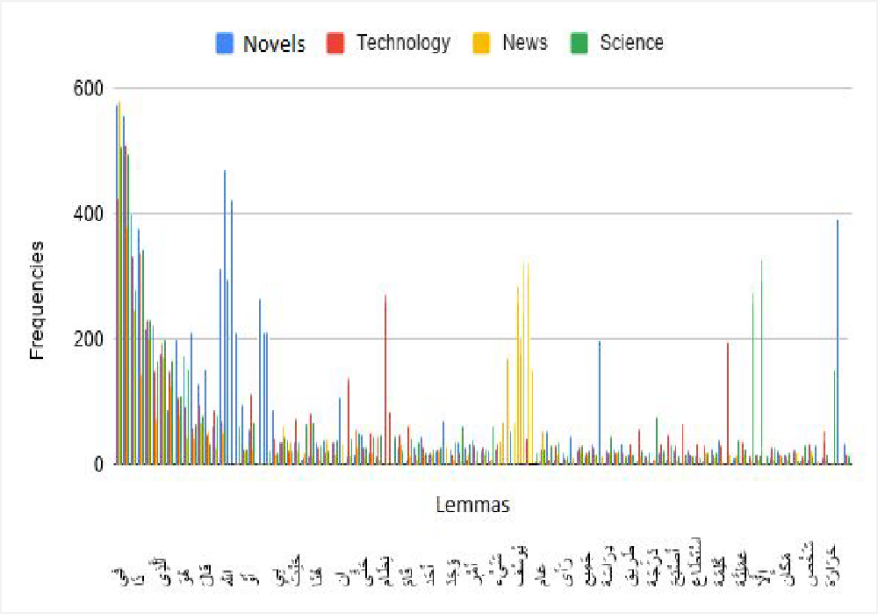


Figure 2. Distribution of the frequent lemmas over the corpus

6. Conclusion

This paper introduced a large-scale speech corpus, which was designed for TTS use. The corpus consists of 15 hours of speech data recorded at 24 kHz sampling rate and its corresponding fully diacritized texts along with the phonetic transcription. To the best of our knowledge, this is the largest Arabic TTS-specific corpus. It's hoped that the release of this corpus will help researchers in the field of speech synthesis. Future work includes evaluating the proposed corpus in training neural end-to-end TTS, and the relationship between the amount of training data and the quality of the produced speech.

References

Al Jazeera (2015). Aljazeera Learn [online]. Available from: http://learning.aljazeera.net/arabic.

Almosallam I., AlKhalifa A., Alghamdi M., Alkanhal M., Alkhairy A. (2013). *SASSC*: *A Standard Arabic Single Speaker Corpus*. In proceedings of 8th ISCA Speech Synthesis Workshop. Barcelona, Spain.

Haamid S., Paulsson N., Choukri K., Shahin M. (2005). *Specifications of the Arabic TTS speakers DB corpus produced within the NEMLAR project*, www.NEMLAR.org.

Halabi N., Wald M. (2016). *Phonetic inventory for an Arabic speech corpus*. In Proc. 10th International Conference on Language Resources and Evaluation, pp. 734-738. Slovenia.

Panayotov V., Chen G., Povey D., Khudanpur S. (2015). *LibriSpeech: an ASR corpus based on public domain audio books,* in Proc. ICASSP, pp. 5206-5210.

Sharief Y., (2017), *End to end Arabic TTS system based on tacotron*, https://github.com/youssefsharief/arabic-tacotron-tts.

Wang Y., Skerry-Ryan R.J., Stanton D., Wu Y., Ron J.W., Navdeep J., Zongheng Y., Ying X., Zhifeng C., Bengio S., Quoc L., Agiomyrgiannakis Y., Clark R., Saurous R. A. (2017). *Tacotron: Towards End-to-End Speech Synthesis.* In Proc. Interspeech, Available from : https://arxiv.org/abs/1703.10135.

Yaseen M., Attia M., Maegaard B., Choukri K., Paulsson N., Haamid S., Krauwer S., Bendahman C., Fersøe H., Rashwan M., Haddad B., Mukbel C., Mouradi A., Al-Kufaishi A., Shahin M., Chenfour N., Ragheb A. (2006). *Building Annotated Written and Spoken Arabic LR 's in NEMLAR Project.* In Proc. 5th LREC-conference, Genoa, Italy.

Zen H., Dang V., Clark R., Zhang Y., Weiss R. J., Ye J., *Zhifeng* C., and Yonghui W. (2019). *LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech*, in arXiv.

DiLiTAL'19, p. 67- 77

Phonemes recognition using formant analysis in the case of consonant vowel transition case "Amazigh language"

Ahmed OUHNINI, Mohammed OUANAN, Brahim AKSASSE

Moulay Ismail University of Meknes, Faculty of Science and Techniques

Department of Computer Sciences, M2I Laboratory, ASIA Team

BP 509 Boutalamine Errachidia, Morocco

ouhnini.ah@gmail.com, mohammed_ouanan@yahoo.fr, baksasse@yahoo.com

Abstract.

Several studies in the development of Automatic Speech Recognition (ASR) systems have been yet presented and well studied for many languages. However, the particularities of the Amazigh Language due to its specific acoustic features such as the emphasis on Amazigh consonants pose many problems for developing its automatic speech recognition system. In this paper, we deal with the spoken recognition's problem in Amazigh Language. We consider here a particular case of the problem, focusing on the vowel-consonant transitions, by only treating the recognition of phoneme level. We propose a derivative algorithm allowing the phoneme separation based on the difference between the energies of both consonants and vowels.

1. Introduction

In order to improve human-machine (HM) communication, Automatic Speech Recognition (ASR) is used to analyze the human voice picked up by a microphone and transcribe it into machine-readable text. ASR is an active research area for many years. A complex process from the production of speech to the determination of spelling, background noise, and sometimes significant differences according to accents and dialects within the same language, the inability to deal with bilingualism, a common situation in some countries, as well as very high computational needs, making real-time recognition very difficult.We consider here a particular case of the problem, focusing on the following transitions between Amazigh language vowels and consonants, and by stopping at the level of phoneme recognition. We believe

that this work will demonstrate the potential of consonant vowel transitions in phoneme recognition by formant analysis.

We propose a derivative algorithm of phoneme separation based on the difference between consonant's and vowel's energy. The formant analysis algorithm focuses on the physical aspect to phonology.

The rest of the paper is considered as follows: In section 2, we present some caracteriscs of the humain voice in general and for the Amazigh lagauge in partcicular. In section 3, we give some mathematical tools used for speech modeling folowed by the proposed method for vowel consonant separation. Furthermore, section 4 presentes some results, and finnally section 5 concludes the paper.

2. Phonetics and phonology elements

2.1. The sound and the Human hearing

The sound is a wave disseminating in a material medium, in the form of small variations pressure. It is perceptible to the human hear, at the point when it's frequency is between 20 Hz, 20 kHz or less. Nevertheless, it is evaluated that the phonetic information is below 10 kHz. When it comes to a phone call, frequency over 3.5 kHz are cut, however the conversation stays comprehensible, while some extremely acute phonemes, similar to [S] are in effect badly rendered.

It is normal in signal processing to just consider spectra of amplitudes and shift phases spectra. This is because of the way that when filtering audio signals, we basically change the sufficiency range on the grounds that the ear is minimal delicate to phase distortions. This legitimizes we just only be keen in the following applications module.

2.2. The Human voice

The human voice results from the interaction between breath and the different phonatory organs collaboration. In the voiced phonems case, the sound is at first delivered by the vibration of the vocal cords. As per the cavities through which it passes, it is handled differently, pharynx and mouth principally. These cavities act as resonators that reinforce for certain phonemes the frequencies corresponding to their resonance frequency. These reinforced frequencies are called formants, and it is them that phonologists try to identify in a spectrogram to recognize pronounced phonemes. It is interesting to note that the nasal cavity attenuates the formants of the so-called nasalized phonemes, and this is why we speak of the phenomenon of anti-resonance.

These phonatory organs are not fixed, their shape varies according to the articulation mode specific to a particular phoneme, and their resonance frequencies, hence the appearance of different formants for each phoneme. In practice, we consider the first three or four formants, the first two being the most related to the mode of articulation. In this study, it was decided to select four of them, designate F1, F2, F3 and F4.

The following figure is a representation of the first three formants (F1, F2, F3) overlayed on the spectrogram of the signal « ⵉⵙⴽⵔ ⵍⵎⵓⵙⵎ: il a pris des vacances » where there is an alternation of voiced and unvoiced sounds. A formant structure is presented, in the voiced case.

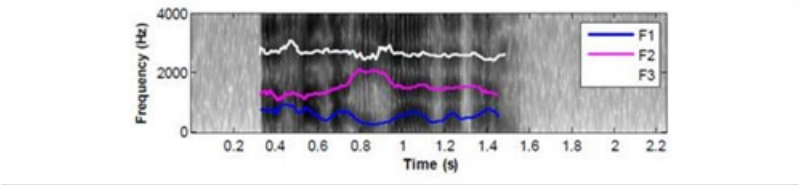


Figure 1. Representation of the first three formants F1, F2 and F3

2.3. Phonetics and phonology

Phonetics is the study of phonemes, the latter being often defined as the smallest distinctive phonetic unit. For example, the opposition of the words bread and bath shows that [p] and [b] are phonemes. They can be classified into several classes and subclasses, the first level being vowels and consonants.

The vowels are sounds that have a fairly long duration, characterized by a free flow and a great constancy of their frequency properties over time: in the absence of too marked prosodic elements, the formants are horizontal on a spectrogram. Vowels are often classified by means of the vowel trapezius, reproduced below, which represents their position in the plane defined by the first two formants and reflects the position of the language during articulation.

The consonants are phonemes that, when articulated, encounter an obstacle (lips vowels for labial, teeth for dental, closing the palate for[k], etc.). They are generally much shorter than vowels and much more variable over time. They can be noisy or sonorous. It is only in the latter case that they present formants.

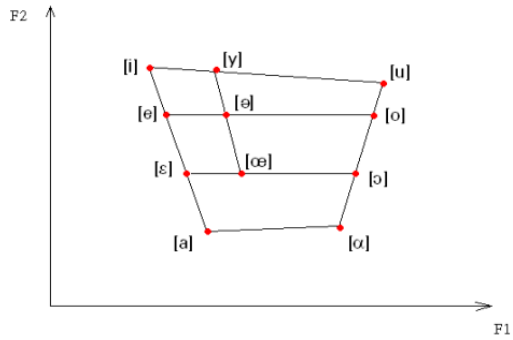


Figure 2. Classification of vowels using the trapezium vowel

Figure 2 represents the spectrogram of the syllable [ⴰⵙⴰ] (asa) and illustrates some of the concepts previously formulated. We can clearly see the horizontality of the formants of the vowel [a] (even if on this unprocessed image, they are difficult to distinguish from other frequency bands "parasites"). The consonant [s] is a very high-pitched noise above 5 kHz, clearly free of formants.

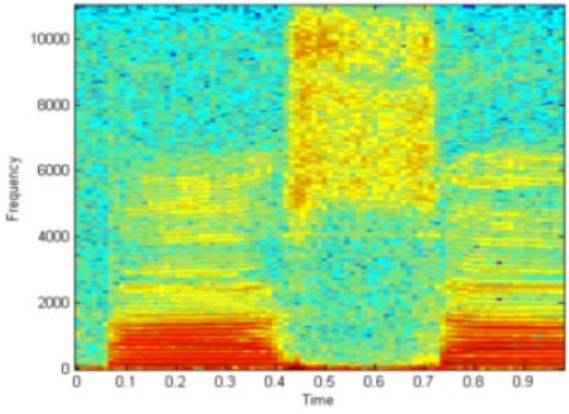


Figure 3. Syllable Spectrogramme [asa]

2.4. The Amazigh language

The Amazigh language, known as Berber or Tamazight, is one of the oldest languages of humanity. Nowadays, it covers the Northern part of Africa which extends from the Red Sea to the Canary Isles, and from the Niger in the Sahara to the Mediterranean Sea. In Morocco, this language is divided, due to historical, geographical and sociolinguistic factors, into three mean regional varieties, depending on the area and the communities: Tarifite in the North, Tamazight in Central Morocco and South-East, and Tachelhite in the South-West and the High Atlas. Eventhough 50% of the Moroccan

population is Amazigh speaker, the Amazigh language was exclusively reserved for familial and informal domains. However in the last decade, thanks to the royal benevolence this language has become institutional, and integrated into the Moroccan education system.

The official moroccan amazigh alphabetical system, known since February 2003 as Tifinaghe-IRCAM, is being implemented in moroccan school curricula and is used in Amazigh historical research. This system contains:

27 consonants	labials	ⵀ, ⵃ, ⵇ
	dentals	ⵜ, ⵟ, ⵠ, ⵟ, ⵍ, ⵏ, ⵕ, ⵖ
	alveolars	ⵙ, ⵣ, ⵚ, ⵥ
	palatals	ⵛ, ⵊ
	velar	ⴽ, ⵅ
	labiovelars	ⴽⵯ, ⵅⵯ
	uvulars	ⵇ, ⵅ, ⵖ
	pharyngeals	ⵄ, ⵃ
	laryngeal	ⴰ
2 semi-consonants	ⵢ and ⵡ	
4 vowels	full vowels	ⴰ, ⵉ, ⵓ
	neutral vowel (or schwa)	ⵔ which has a rather special status in Amazigh phonology.

Table 1. Amazigh language consonants, semi-consonants and vowels.

The spelling of the words, in latin characters, very close to phonetic transcription, accurately reflects the pronunciation, including vowels and twinned.

Example: /illa/—> « illa » —> « il est », « il existe ».

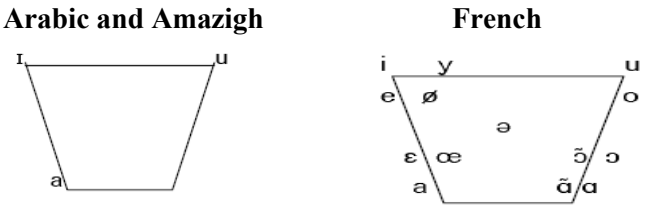


Figure 4. Comparison of the Arabic, the Amazigh and the French vocal systems

The Amazigh language differs in pronunciation from region to region. In our work, we considered the central and south-eastern region of Morocco.

3. Methodology

3.1. Fourier analysis

The Fourier transform allows a time-frequency analysis to be performed with sufficient resolution for the voice signal, quasi-stationary over intervals in the range of 10 to 100 ms.

3.1.1. Fourier Transform

We are working in the set of square integrable functions $L^2(\mathbb{R})$, which is a pre-Hilbert, and we consider the orthogonal family of sine $e_\omega = t \to e^{i\omega t} / \omega \in \mathbb{R}$ (of course, physically, we limit ourselves to positive pulses). The Fourier transform F (or TF(f)) of some Function permits a projection onto the vector space spanned by the sinusoids first by preparing the part of the projection on every sinusoid given by the scalar produce, and not overlooking the complex conjugation in relation to the second place:

$$F(\omega) =< f \mid \mathrm{e}_\omega >= \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt \qquad (1)$$

In our study, this transform will never be used to reconstruct the signal, but to a more qualitative way to examine the contributions of every frequency range in the speech signal by studying the spectrogram, which is in fact a three-dimensional representation indicating the magnitude (modulus of F(ω)) in the time-frequency plane. (According to Ohm's law, the ear is not sensitive to the phase of the auditory signal.)

However, this continuous representation can't be digitally processed and then must be adapted to the discrete case of the digital calculation.

3.1.2. Discrete Fourier Transform

The signal is represented by samples taken uniformly over time:

$$\{f(n)/n \in [0, N-1]\}$$

We make of them a signal of period N in order to avoid edge effects. The discrete Fourier transform is then given by:

$$F(k) = \sum_{n=0}^{N-1} f(n) e^{\frac{-2i\pi kn}{N}} \qquad (2)$$

The observation of frequency is proportional to $\frac{k}{N}$ factor for $k \in [0, \frac{N}{2}]$: If f_e is the sampling frequency, then $f = f_e \cdot \frac{k}{N}$. It should be noted that this transform is redundant, since the information obtained is concentrated in half of interval $[0, N-1]$. Indeed:

$$\forall k \in Z \; F(N-k) = \sum_{n=0}^{N-1} f(n) e^{\frac{2i\pi kn}{N}} e^{-2i\pi n} = \sum_{n=0}^{N-1} f(n) e^{\frac{2i\pi kn}{N}} = \overline{F(k)} \qquad (3)$$

Hence the module is the same, the phase being simply of opposite sign. If we want to analyze the signal over a range of 10 kHz, so a sampling frequency of 20 kHz should be used. In practice, we implement the Fast Fourier Transform (FFT) which has a complexity of O(Nlog2(N)) instead of $O(N^2)$ by direct calculation and which takes advantage of this redundancy to optimize the calculation.

3.2. Fourier transform and convolution

In continuous representation, the convolution of two functions is defined by:

$$\forall x \in R, (f * g)(x) = \int_{-\infty}^{+\infty} f(t) g(x-t) dt \qquad (4)$$

By changing the variable to $u = x - t$, the convolution operator is commutative. It is also associative:

$$((f * g) * h)(x) = \int_{-\infty}^{+\infty} f * g(x-y) h(y) dy = \int_{-\infty-\infty}^{+\infty+\infty} f(x-y-t) g(t) h(y) dt dy$$

$$= \int_{-\infty}^{+\infty} g(t) \left(\int_{-\infty}^{+\infty} f(x-t-y) h(y) dy \right) dt = \int_{-\infty}^{+\infty} g(t) (f * h)(x-t) dt$$

$$= (g * (f * h))(x) \qquad (5)$$

and therefore by commutativity we have:

$$(f * g) * h = (f * h) * g \qquad (6)$$

The Fourier transform of the convolution product of two functions is the usual product of Fourier transforms

$$TF(f * g) = F \times G \qquad (7)$$

This outcome is also valid in discrete representation. It will be utilized in formant analysis consequence of the modelling adopted for the voice signal.

3.3. Windowing problem

In practice, not only the signal is discreet, but also the observation time 2τ is over. We therefore observe the convolution of signal through windowing function:

$$\Pi : t \to 1 \;\; si\, t \in [-\tau, \tau]$$
$$0 \;\; sinon$$

According to (7), the Fourier transform of s ∗ Π is the product of the transforms, that of the door function being:

$$\int_{-\infty}^{+\infty} \Pi(t) e^{-i\omega t} dt = \int_{-\tau}^{\tau} e^{-i\omega t} dt = 2\tau \,\mathrm{sinc}(\omega\tau) \qquad (8)$$

The central lobe of this cardinal sinus is 2/τ wide. Consequently, when the observation time tends towards zero, the spectrum widens. To solve this problem, windows with energy concentrated in zero are used, which limits this phenomenon. In our work, we used the Hamming window.

3.4. Formant analysis

3.4.1. Acoustic-phonetic decoding problem

Since the voice signal is continuous, it is difficult to recognize in the recorded signal the different linguistic elements that are the words, syllables and phonemes. That problem is called acoustic-phonetic decoding. Here, we have utilized a technique for identify the transitions between vowels and consonants. The syllables for that occasion are available to qualified. They are in the great majority in some languages, such as the Polynesian or Japanese languages to a lesser extent, but this is not the case in the languages where two consonants can often follow each other, or worse, in Slavic languages, such as Czech. It is therefore only a partial solution to the issue, particularly as the words to be analyzed must to have been previously cleared of initial and final silences.

3.4.2. Consonant-vowel and vowel-consonant transitions detection

To eliminate as much background noise as possible, digital filtering is first used. A preliminary analysis having shown that, under the conditions of the experiment, the background noise is mainly located below 100 Hz (and mainly at 50 Hz, the frequency at which turn the fans of computers), we firstly invokes the function that calculate the FFT of the signal, cuts harmonics below 100 Hz (taking into account redundancy and periodicity) and returns the inverse FFT.

The proposed algorithm is based on the fact that the sound power of consonants is generally much lower than that of vowels. We start by calculating the root mean square at each point over a wide enough window to be larger than the possible period of the phoneme (which at this stage is clearly not known). The signal is then derived, after being smoothed to presence peaks due to simple fluctuations, and then the derivative itself is smoothed in order to be simpler to process. From that point, we search for (φ-1) most important peaks in absolute value, which correspond to the most brutal variations (excluding the edge, on which outlier phenomena occur).

If this peak is negative, the power has dropped sharply, so we are in the presence of a Vowel-Consonant (VC) transition, otherwise it is a Consonant-Vowel (CV) transition.

The figure 4 below illustrates our method for calculating the acoustic power of the syllable [ara]. Forever, the figure 5 shows the temporal derivative of the acoustic power given in figure 4.

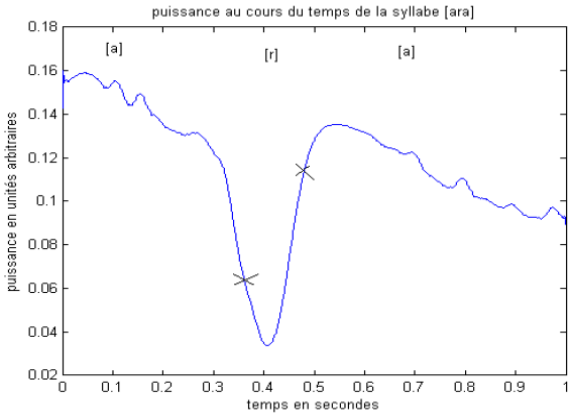


Figure 5. Acoustic power [ara]

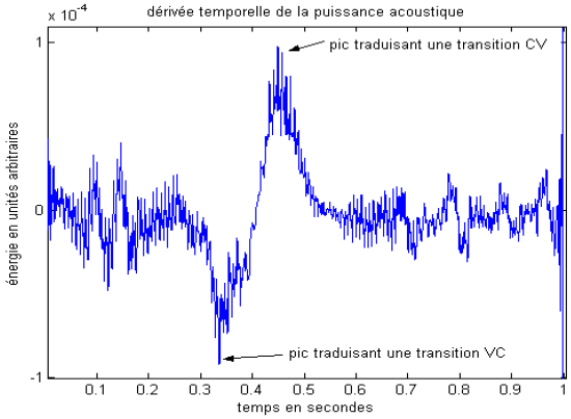


Figure 6. Derivative of acoustic power [ara]

It is empirically estimated (after a series of experiments) that the transition is located at a moment in the vicinity of the peak where, compared to the previous phoneme, the signal has lost or gained a power equal to 66% of the absolute difference in power between the two phonemes. On the graph, the transitions are indicated by crosses.

4. Processing results

The phoneme separation algorithm works very well as long as:

- The beginning and end silences have been deleted;
- The speaker does not blow on the microphone during recording, which brings the signal to saturation level and results in very large peaks that the algorithm takes for a transition;
- We limit ourselves to a maximum of six phonemes (beyond that, the algorithm no longer succeeds in distinguishing power variations between two phonemes and fluctuations);
- This system achieves a performance rate of 75%;
- Transitions are more important for vowels that last much longer than consonants. This explains the difficulty of detecting transitions.

5. Conclusion

In this paper we have considered the problem of automatic speech recognition for the amazigh language. We have partcicullary tried to separate the consonants and vowels of Amazighe voice signal. It is currently possible to make an Amazigh corpus to join those really accessible. This methodology gives a portal to a social and economic development of Moroccan Amazigh society.

This approach to speech recognition has allowed us to identify certain characteristics phonetics and to implement solutions to exploit them. However, it can be improved in the context of a future work. If the software, upon arrival, does not allow to identify precisely each phoneme, it highlights the proximity of a good many of them. To obtain better results, several ways are possible: a better determination of the problem parameters (emphasis coefficient, frequency cut cepstrum and so on.) as well as a more precise analysis of trajectories and formant transitions and a better consideration account of prosody in the speech.

Bibliography

Dave. N. (2013). Feature Extraction Methods LPC, PLP and MFCC in Speech Recognition, International Journal for Advance Research in Engineering and Technology, Vol. 1, Issue 6.

Ghazi E. (2011). Système de reconnaissance automatique de la parole amazighe, Revue Méditerranéenne Télécommunications, Vol. 1.

Abdenbi A., Ataa Allah F., Benayad N. (2014). Vers un système de reconnaissance automatique de la parole en amazighe basé sur les transformations orthogonales paramétrables, Asinag, n° 9, pp. 133-145.

Juang B. H., Rabiner L. R. (2004). Automatic Speech Recognition – A Brief History of the Technology Development, Georgia Institute of Technology, Atlanta and Rutgers University and the University of California, Santa Barbara.

Gendrot C. (2004). Analyses formantiques automatiques de voyelles orales : évidence de la réduction vocalique en langues française et allemande, MIDL.

Hosni H., Sakka Z., Kachouri A., Samet M. (2009). Étude de la Paramétrisation RASTA PLP en vue de la Reconnaissance Automatique de la Parole Arabe, 5th International Conference: Sciences of Electronic Technologies of Information and Telecommunications, Tunisia.

Agrawal M., Raikwar T. (2016). Speech Recognition Using Signal Processing Techniques. International Journal of Engineering and Innovative Technology, Vol. 5, n° 8.

Samudravijaya K. Modeling Natural Language for Automatic Speech Recognition, Tata Institue o Fundamental Research Homi Bhabha Road, Mumbai 400005, India.

Lecouteux B. (2008). Reconnaissance Automatique de la Parole, Thèse Université d'Avignon et des Pays de Vaucluse.

Dufour R. (2010). Transcription automatique da la parole spontanée, Thèse de l'Université du Maine.

Ataa Allah F., Boulaknadel, S. (2010). Amazigh Search Engine: Tifinaghe Character Based Approach. International Conference on Information and Knowledge Engineering.

DiLiTAL'19, p. 79 - 89

Dictionnaire général de la langue amazighe informatisé

Fadoua ATAA ALLAH, Jamal FRAIN, Siham BOULAKNADEL

Centre des Etudes Informatiques, des Systèmes d'Information et de Communication, Institut Royal de la Culture Amazighe, Rabat, Maroc

{ataaallah,boulaknadel,frain}@ircam.ma

Résumé.

Dans la perspective de la préservation et la revitalisation de la langue amazighe, son informatisation est une démarche qui s'impose à cette ère du numérique. Le processus de la concrétisation de cette démarche décline et structure un ensemble de ressources élémentaires. Le dictionnaire numérique, dont fait partie, est l'un des piliers de ce processus. Dans ce contexte, l'Institut Royal de la Culture Amazighe a réalisé le Dictionnaire Général de la Langue Amazighe Informatisé. Le présent article introduit ce dictionnaire et expose sa conception. En outre, il introduit l'évolution des dictionnaires et présente un ensemble de types de dictionnaire.

1. Introduction

De nos jours, grâce aux avancements technologiques, l'informatique n'est plus un objet d'étude restreint sur des personnes émérites, mais elle est devenue un outil d'usage qui séduit différentes catégories d'âge et socioprofessionnelles. Ainsi, dans l'objectif de faire bénéficier la langue et la culture amazighes des avancées de cette technologie, il est indispensable de les doter d'un ensemble de ressources et d'outils facilitant leur usage et assurant leur diffusion (Ataa Allah et Boulaknadel, 2012 ; Ataa Allah et Boulaknadel, 2014). Dans cette optique, le Centre des Etudes Informatiques, des Systèmes d'Information et de Communication (CEISIC), relevant de l'Institut Royal de la Culture Amazighe (IRCAM), en collaboration avec le Centre d'Aménagement Linguistique (CAL) a entrepris la production d'un dictionnaire numérique, qui s'intitule « Dictionnaire Général de la Langue Amazighe Informatisé » (DGLAI).

Selon la définition du trésor de la langue française[1], un dictionnaire est un « Recueil des mots d'une langue ou d'un domaine de l'activité humaine, réunis selon une nomenclature d'importance variable et présentés généralement par ordre alphabétique, fournissant sur chaque mot un certain nombre d'informations relatives à son sens et à son emploi et destiné à un public défini. ». À partir de cette définition, l'objectif de la confection de dictionnaire se ramifie en plusieurs sous-objectifs, qui interpelle une réflexion sereine, afin de réussir l'élaboration d'un dictionnaire particulier. Parmi les questions auxquelles il faut répondre, c'est qu'il est le public visé et qu'est ce qui lui apportera ce dictionnaire ? En déterminant ces objectifs, la nomenclature du dictionnaire se trace et sa structure s'engrave.

Le reste de cet article est organisé comme suit : la deuxième section introduit brièvement l'évolution des dictionnaires, depuis leur naissance jusqu'à cette ère du numérique. La troisième section est consacrée à la catégorisation de types de dictionnaires. La quatrième section présente la méthodologie de conception du DGLAI et sa structure. Tandis que la quatrième expose une conclusion et décline les perspectives de ce travail.

2. Evolution des dictionnaires

Le dictionnaire est un objet qui illustre la conception du discours d'une époque et un miroir qui reflète la richesse d'une culture. Il joue un rôle important dans la consignation des éléments culturels méconnus par des personnes qui ne partagent pas la même culture. En outre, il permet la transmission de l'usage de mots et d'expressions à charge culturelle aux apprenants, comme le précise Pruvost (1999) « un grand nombre de mots et d'expressions dépendent d'un implicite culturel déterminant pour la pleine compréhension du locuteur natif, un implicite auquel il faut initier l'élève en cours d'apprentissage. Une culture spécifique, mobilisée et actualisée dans les mots d'une langue, imprègne ainsi fortement la communication d'un discours second dont le rôle est souvent décisif ».

L'histoire des dictionnaires remonte à la seconde moitié du 4^{e} millénaire av. J.-C., où les Sumériens ont inventé des listes lexicales, afin d'aider à l'apprentissage du sumérien et de l'écriture cunéiforme. Ces listes sont constituées des énumérations de signes (syllabaires) et de mots (vocabulaires) présentés en colonnes et classés selon des critères thématiques, sémantiques ou graphiques (Michel, 2001). A la fin du 3^{e} millénaire av. J.-C., le sumérien est devenu une langue morte. Néanmoins dans l'intérêt de l'apprentissage, les scribes akkadiens ont continué à

[1] http://atilf.atilf.fr/

recopier, compléter et développer ces listes qui sont restées monolingues jusqu'au milieu du 2e millénaire. Par la suite, ces listes lexicales sont devenues bilingues suméro-akkadiennes. Elles étaient découvertes sous forme de tablettes à Ebla en Syrie moderne. Elles contiennent les mots sumériens énumérés dans la colonne de gauche et leur traduction en akkadien dans la colonne de droite.

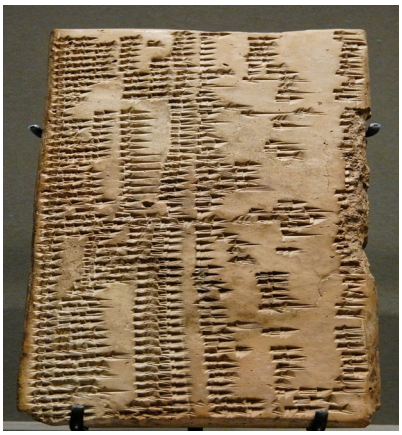


Figure 1. 16e tablette des listes sumério-akkadiennes (Musée du Louvre)

Depuis, les dictionnaires ont connu une évolution continuelle liée à plusieurs niveaux, à savoir : la conception du dictionnaire, sa fabrication, sa réception et sa consultation. Toutefois, la révolution humaine pourra enchaîner d'autres métamorphoses dans l'avenir. Les métamorphoses perceptibles dans la conception des dictionnaires sont fortement liées à leur fabrication et notamment à la nature du support d'écriture utilisé et affectent bien évidemment les métamorphoses de réception et de consultation.

Le passage de l'écriture sur des tablettes en argile puis en bois à l'écriture sur papier de papyrus ou d'amate, ensuite sur des parchemins et l'assemblage sous codex, assurant l'attachement des supports entre eux en s'assimilant à un livre composé de feuilles, a ouvert aux dictionnaires une ère moderne. Cette ère leur a garanti une longue vie sur papier en gros volume, puis imprimé en petit nombre pour arriver jusqu'au dictionnaire de poche. Facilitant, ainsi, la consultation d'une entrée référencée par une table des matières ou par un ordre alphabétique.

Cette ère moderne a été couronnée par l'ère du numérique, offrant aux dictionnaires une considérable évolution par le passage de l'argile durcie à la plaque de métal et ses scintillements, passant paradoxalement du matériel plus dur, la roche et l'argile, au fragile, le papier, au plus fragile, le support dématérialisé, permettant la transmission d'un bout à l'autre de la planète.

Néanmoins, l'évolution des dictionnaires, au cours de l'ère du numérique, a connue aussi une progression allant de versions sur cédérom, dans les années 80s, à des versions sous forme d'applications mobiles, de nos jours, passant par des versions consultables intégralement sur Internet et assorties de moteurs de recherche, dans les années 2000s.

Grâce à la révolution numérique, le contenu d'un dictionnaire peut être consultable, désormais, à travers quatre approches différentes selon le support choisi : papier, cédérom, Internet ou application mobile, tout en permettant une panoplie de manières d'organisation de ce contenu.

3. Catégories de dictionnaires

Selon les objectifs visés par la confection d'un dictionnaire, sa nomenclature et sa structure peuvent varier. D'ailleurs selon les travaux Kasar (2008), les dictionnaires peuvent être classifiés en six grandes catégories, selon six critères fondamentaux, à savoir :

1. le mode de présentation ;
2. le matériau de la microstructure[2] du dictionnaire ;
3. le nombre et le statut des langue(s) étudiée(s) ;
4. les propriétés des informations fournies ;
5. l'extension de la macrostructure ;
6. l'organisation de la macrostructure.

Ainsi, de chaque catégorie liée à l'un de ces critères se déclinent divers types de dictionnaires, qui peuvent être énumérés comme suit :

Catégorie 1 : dictionnaires papier, dictionnaires électroniques, dictionnaires numériques ;

Catégorie 2 : dictionnaires (proprement dits), dictionnaires illustrés ;

Catégorie 3 : dictionnaires monolingues, dictionnaires plurilingues ;

Catégorie 4 : dictionnaires de mots ou de langue, dictionnaires de choses ou encyclopédiques ;

Catégorie 5 : dictionnaires extensifs, dictionnaires restrictifs ;

Catégorie 6 : dictionnaires alphabétiques, dictionnaires thématiques.

[2] Selon Hausmann (1988), un dictionnaire est une imbrication d'une structure binaire de la macrostructure et de la microstructure. La macrostructure est la structure d'accès principale du dictionnaire. Elle est constituée par l'ensemble des entrées formant la nomenclature. Cependant, la microstructure présente l'ensemble des informations présentées dans chaque entrée.

3.1. Catégorie liée au mode de présentation

Jusqu'aux dernières décennies, il existait un seul mode de présentation qui est le dictionnaire papier. Cependant, les avancées techniques ont établi de nouvelles formes de mode de présentation des dictionnaires, il s'agit des dictionnaires électroniques et des dictionnaires numériques qui se consultent sur écran.

Un dictionnaire électronique est un appareil électronique servant de dictionnaire. Le premier dictionnaire électronique a été lancé en 1981 par Casio. Il est bilingue (anglais-japonais) de 2020 mots et expressions idiomatiques, d'une épaisseur de 8 mm exactement et pesant 53 g.

Un dictionnaire numérique peut être comparé à Janus au double visage, avec un visage tourné vers la linguistique-informatique, qui reste son principal domaine d'application, et l'autre tourné vers la lexicographie, de laquelle il naît comme relecture particulière des buts du dictionnaire papier. Même transcrit sur support informatique, il est destiné à l'usage humain. Il est orienté vers la définition des mots et la signification de leurs emplois et partage avec le dictionnaire traditionnel les finalités de catégorisation et de description morpho-grammaticale du lexique. Cependant, les dictionnaires numériques construits pour la machine, en outre du listage des entrées utilisées dans une langue donnée, ils focalisent sur la description formelle des objets de la langue et leur classification systématique. Ils représentent des ensembles très spécifiques de données qui sont élaborés et constituants un système informatif dans sa globalité.

3.2. Catégorie liée à la microstructure

La majorité des dictionnaires conçus en langue(s) naturelle(s) ce sont des dictionnaires proprement dits, dont la macrostructure comme la microstructure sont tissées en matière linguistique. Toutefois, à l'égard des nouvelles conceptualisations pédagogiques et lexicographiques, de nos jours, il existe des dictionnaires illustrés en images. Ces derniers présentent l'avantage de passer aussitôt du signifiant au signifié en évitant l'intermédiaire d'une langue seconde. Contrairement aux dictionnaires proprement dits, les dictionnaires illustrés sont organisés généralement par thèmes suivant la nécessité.

3.3. Catégorie liée aux nombre et statut de langues

Au sein de cette catégorie, deux grandes classes de dictionnaires se distinguent et se varient dans leur groupe. Il s'agit de dictionnaires monolingues, lorsqu'ils traitent une seule langue, et de dictionnaires plurilingues, quand ils traitent deux ou plusieurs systèmes linguistiques. Les

dictionnaires plurilingues se distinguent entre eux par : un dictionnaire homoglosse et un dictionnaire hétéroglosse.

Le dictionnaire homoglosse est un recueil lexicographique où la langue d'entrée est soit un parler ou un dialecte de la langue de sortie, qui est la langue commune, soit la langue ancienne de la langue de sortie, qui est la langue moderne. Dans le premier cas, il s'agit de dictionnaire de dialectes ; cependant, le deuxième cas concerne le dictionnaire de l'ancienne langue.

Les dictionnaires hétéroglosses, eux aussi, se divisent en deux : il s'agit de dictionnaires bilingues, s'ils traitent deux langues, ou de dictionnaires multilingues s'il y a plus de deux langues dans le dictionnaire. Les dictionnaires multilingues, à leur tour, se regroupent en deux : dictionnaires de langues anciennes et dictionnaires de langues vivantes.

3.4. Catégorie liée aux propriétés de l'information

Dans cette catégorie, deux types à distinguer : dictionnaires de mots ou dictionnaires de langue et dictionnaires de choses ou dictionnaires encyclopédiques.

Le dictionnaire de mots ou le dictionnaire de langue présente des informations sur les signes linguistiques alors que le dictionnaire de choses ou le dictionnaire encyclopédique expose des informations sur les choses. Les dictionnaires de mots se regroupent en deux : dictionnaires généraux et dictionnaires spécialisés qui présente diverses informations sur les mots ; par exemple dictionnaires de racines, dictionnaires de verbes, dictionnaires d'orthographe ou dictionnaires analogiques.

3.5. Catégorie liée à l'extension de la macrostructure

Les dictionnaires, de cette catégorie, se divisent en deux groupes : il s'agit, d'une part, des dictionnaires extensifs, qui considèrent toutes les unités d'une langue y compris les mots anciens ou dialectaux, ou les divers registres de langue. D'autre part, il y a les dictionnaires restrictifs qui limitent leur nomenclature selon un but ou un public visé : soit ce sont des dictionnaires restrictifs spécialisés qui sont des vocabulaires techniques, des terminologies de spécialités et de professions, soit il s'agit de dictionnaires d'usage qui peuvent être de trois natures différentes : descriptifs, prescriptifs ou correctifs.

3.6. Catégorie liée à l'organisation de la macrostructure

Relativement à la structure de la macrostructure, les dictionnaires se présentent en deux types : dictionnaires alphabétiques et dictionnaires thématiques. Les dictionnaires alphabétiques introduisent les entrées selon

l'ordre alphabétique de la langue traitée tandis que dans les dictionnaires thématiques, le traitement des unités est fait par thèmes.

4. Du Dictionnaire Général de la Langue Amazighe à sa version informatisée

L'informatisation d'un dictionnaire permet, avec les avancées des techniques informatiques, de percevoir des parcours nécessairement novateurs ainsi que des emplois nouveaux concernant la lecture et la recherche dans les textes. Ces enjeux ont été promptement dégagés et, avant même la fin de sa publication papier, l'informatisation du Dictionnaire Général de la Langue Amazighe : Amazighe-Français-Arabe (DGLA) a été envisagée. Dans cette informatisation, trois éléments sont survenus :

- une collaboration scientifique avec des linguistes relevant du Centre d'Aménagement Linguistique de l'IRCAM ;
- un soutien sans défaut de l'IRCAM qui attendait beaucoup d'une telle informatisation, ressource de base pour des recherches futures ;
- le travail réalisé par le Centre des Etudes Informatiques, des Systèmes d'Information et de Communication, qui sut affirmer que la faisabilité de la rétroconversion du DGLA pouvait être réalisée.

Le DGLAI est un dictionnaire numérique, destiné à l'usage humain. Il est fondé sur des spécificités liées à son contenu, notamment à sa nomenclature, qui est d'environ 13 674 entrées relevant de l'usage et représentant toutes les variétés marocaines, et sa liste d'objets métatextuels, sur lesquels repose la structuration de l'article du dictionnaire : vedettes, prononciation en Alphabet Phonétique International (API), variantes, codes grammaticaux, sens, équivalents dans autres langues et emploi.

4.1. Présentation du DGLAI

Le DGLAI est un dictionnaire numérique, qui se présente à la fois comme une base lexicologique et une base de connaissances dont l'accessibilité est immédiate via l'internet et sur différents supports, à savoir : ordinateur, tablette et téléphone intelligent. Son site Web[3] est menu d'un guide d'utilisation et il est doté d'une interface de consultation facilitant la recherche, à travers un module d'auto-complétion, et simplifiant la saisie en tifinaghe ou en transcription latine par le biais d'un clavier virtuel.

[3] https://tal.ircam.ma/dglai/

DGLAI
Accueil | Contact | Télécharger Dictionnaire
Dictionnaire Général de la Langue Amazighe Informatisé
Amazighe – Français – Arabe
DGLAI en chiffres
13 674 entrées en caractères Tifinaghe suivis d'une transcription en A.P.I
40 549 équivalents pour illustrer les sens des différentes entrées
Les auteurs : Meftaha Ameur, Khalid Ansar, Abdallah Boumalk, Noura El Azrak, Rachid Laabdelaoui, Hamid Souifi, Lahoucine Amouzay, Aïcha Bouhjar, Abdesslam Boumisser.
Les réalisateurs : Fadoua Ataa Allah, Siham Boulaknadel, Jamal Frain.
Présentation
Le DGLAI est la version électronique du Dictionnaire Général de la Langue Amazighe. Il est le fruit d'une collaboration entre le Centre d'Aménagement Linguistique et le Centre des Etudes Informatiques, des Systèmes d'Information et de Communication. Il vient répondre aux attentes du public en mettant à la disposition de l'usager potentiel un utilitaire lexicographique. Il est réalisé en une version trilingue : amazighe-français-arabe.
Le DGLAI se distingue par la finesse de la structuration des données. Une interface simple et conviviale qui offre deux niveaux de consultation : recherche simple et assistée. Ainsi, à travers le DGLAI, l'usager peut bénéficier d'une recherche d'un mot en orthographe approximative ou d'une recherche par les différentes acceptions et langues. En outre, le DGLAI offre la possibilité de consulter un renvoi ou de passer à une autre entrée par un simple clic de souris.
Entrer dans le DGLAI
Copyright@2018 IRCAM

Figure 2. Page d'accueil du DGLAI

Le DGLAI se différencie des autres dictionnaires numériques par la finesse de la structuration des données en entrées interrogeables selon divers critères, et par une interface simple et conviviale qui offre la recherche en trois langues (l'amazighe, l'arabe et le français) et en trois manières :

- Une recherche approximative : sous forme de proposition de consultation de mots en affichant une liste de mots voisins soit par leur orthographe (proximité morphologique) ou par proximité alphabétique (mots commençant par…), et ce grâce au module d'auto-complétion.

- Une recherche simple : sous sa forme habituelle figurant en entrée en tifinaghe ou sous une forme transcrite en API dans le dictionnaire.

- Une recherche avancée :
  - sous une forme fléchie, c'est-à-dire son pluriel, son état d'annexion ou une forme conjuguée selon les trois aspects (accompli, accompli négatif, inaccompli), quand la langue de recherche est l'amazighe ;
  - sous une forme d'emploi ;
  - sous une forme exacte ou contenant une partie du mot saisi.

Le résultat de la recherche est organisé différemment selon le nombre d'entrées liées au mot désigné par la recherche. Lorsque le résultat est constitué de plus d'une entrée, ces dernières sont affichées en deux listes :

Figure 3. Page de recherche du DGLAI

▪ la première liste contient le mot tel qu'il est saisi avec sa variante s'elle existe, sa prononciation, sa nature grammaticale, suivi de sa morphologie, ses équivalents en langues française et arabe structurés en sens, ainsi que son mode d'emploi ;

▪ la seconde liste appelée « Recherche en relation » affiche les différentes entrées de ce mot, avec sa ou ses catégorie(s) grammaticale(s), suivi de son équivalent en langue française.

4.2. Structure du DGLAI

La conception de la base de données du DGLAI s'est fondée sur un ensemble de critères, y compris :

▪ le DGLA est un dictionnaire de la langue amazighe, qui prend en considération toutes les variantes amazighes ;

▪ sa macrostructure est basée sur les lexies amazighes ;

▪ il présente les équivalents de chaque entrée amazighe en langues arabe et française.

A la différence des travaux précédents d'élaboration de dictionnaires numériques de la langue amazighe (Nejme *et al.*,2016 ; Taghbalout *et al.*, 2017) destinés pour la machine, ce travail envisage l'usage humain.

La microstructure du DGLAI est composée d'une suite de paires attributs/valeur. Chaque lexie est décrite à travers une représentation en un graphe d'attributs, qui contiennent chacun une partie de l'information relative à la lexie.

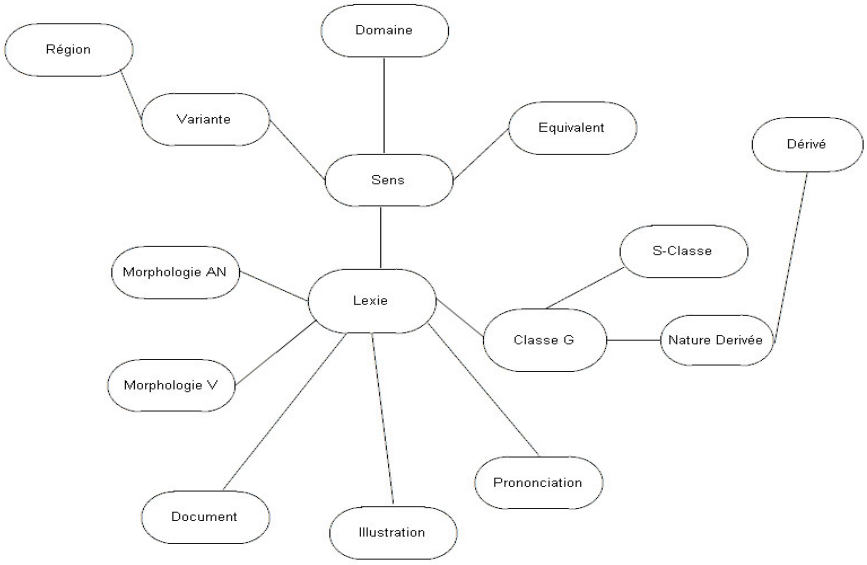


Figure 4. Structure de la base de connaissances du DGLAI

Cette représentation est un réseau plutôt qu'une forme linéaire. L'entrée lexicale pointe sur des attributs linguistiques et d'autres extralinguistiques : l'attribut illustration ; l'attribut prononciation, offrant la possibilité d'avoir des enregistrements sonores ; et l'attribut document, qui permet d'utiliser des corpus pour dégager des exemples authentiques de la langue.

Les attributs linguistiques incluent l'attribut de partie du discours (classe grammaticale) ; l'attribut des différentes variantes possibles d'une entrée ; les attributs morphologiques ; l'attribut du sens ; lequel est en relation avec l'attribut des équivalents en d'autres langues. Spécifiquement, cela signifie que la description des équivalents se décline du sens.

Chaque entrée a un identifiant, qui permet de situer cette entrée dans le système de renvois multiples de la base de données. La structure du dictionnaire ainsi conçu pourrait aussi être exploitable aussi pour décrire la nature de dérivation d'une entrée lexicale.

5. Conclusion

Certaines des ressources linguistiques constituent un patrimoine important, dont le souhait de les faire connaître et de les rendre accessibles à tous. Le Dictionnaire Général de la Langue Amazighe, conçu au départ sous sa version papier est devenu grâce à son informatisation et sa mise à la disposition de tous, une ressource incontournable sur la langue amazighe.

Références

Ameur M., Ansar, K., Boumalk A., El Azrak N., Laabdelaoui R. (2017). Dictionnaire général de la langue amazighe. Publications de IRCAM, Rabat.

Ataa Allah F., Boulaknadel, S. (2012). Toward computational processing of less resourced languages: primarily experiments for Moroccan Amazigh language. Theory and applications for advanced text mining, Rijeka: Intech.

Ataa Allah F., Boulaknadel, S. (2014). La promotion de l'amazighe à la lumière des technologies de l'information et de communication. Asinag, n° 9, pp. 33-48, IRCAM publications.

Hausmann F. J. (1988). De quoi se compose l'article du dictionnaire de langue ? L'importance du sous-adressage. Actes de 3rd EURALEX International Congress, pp. 59-664-9 septembre 1988, Budapest, Hongrie.

Kasar S.Ö. (2008). Pour une typologie moderne des dictionnaires. Synergies Turquie, n° 1, pp. 55-64

Michel C. (2001). Listes lexicales, dans F. Joannès, assisté de C. Michel, Dictionnaire de la civilisation mésopotamienne, Paris : Robert Laffont, Coll. BOUQUINS, pp. 475-476.

Nejme F.Z., Boulaknadel S., Aboutajdine D. (2016). AmAMorph: Finite State Morphological Analyzer for Amazighe. Journal of Computing and Information Technology, 24(1):91-110.

Pruvost J. (1999). Lexique et vocabulaire : une dynamique d'apprentissage, *Vocabulaires et dictionnaires en FLM et FLE*, Études de linguistique appliquée, n°116, Paris, Didier érudition, pp. 395-421.

Taghbalout I., Ataa Allah F., El Marraki M. (2017). Pivot-based Multilingual Dictionary Model for Under-resourced Languages, International Journal of Applied Engineering Research, 12(20):10402-10410.

DiLiTAL'19, p. 91 - 98

Feature selection from an expanded Moroccan opinionated dataset

Ibtissam TOUAHRI, Azzeddine MAZROUI

Department of computer science, Faculty of Sciences, University Mohamed First, Oujda

{ibtissamtouahri555,azze.mazroui}@gmail.com

Abstract.

Sentiment analysis is a task performed over the extracted opinionated information, in order to describe opinion orientation where the information carries a set of feelings that may be positive, negative, mixed or neutral. Thus, sentiment analysis identifies the polarity of a specific opinion based on natural language processing and machine learning techniques. Training a model on a set of extracted features doesn't address the complexity problem. Hence, a pretreatment task that aims to select features is indispensable, in order to reduce training time by filtering redundant and fewer information features. In this paper, we describe the basic knowledge of our system by identifying the employed linguistic resources, the collection manner, the annotation process, and the associated statistics. Then, we describe the employed input features to construct our model and the manners to optimize training time.

1. Introduction

Knowledge extraction from opinionated data aims to perform a deep analysis of a set of opinions in order to make decisions. New technologies facilitate analysis process and reduce human effort when performing a sentiment analysis task.

Sentiment analysis is an active area in natural language processing and it is based on linguistic characteristics and machine learning techniques. Building a sentiment analysis system based on a supervised approach requires training a model on a set of extracted features. The manners of extraction differ and may generate redundant or noisy features. Hence, feature selection is a crucial preprocessing step that reduces training time by selecting the most significant features. Moreover, it may allow performance improvement.

Features number decrease allows a better understanding of the generated model performance.

Feature selection is a pre-processing step that follows feature extraction and aims to optimize features by reducing their dimensionality. Thus, it improves execution time and diminishes storage requirements. Afterward, the selected features must be compared to general ones in terms of size, execution time, noise persistence, performance, model readability and result interpretation. Among feature selection methods, we find filter-based algorithms that specify selection metrics. It ranks features by assigning weighs, it also offers the possibility of specific features selection by specifying the number of features. The feature vector may be selected by performing other approaches like those based on term occurrence.

In the following, we will present some previous works, in section 2. In section 3, we describe the linguistic resources of our system. We detail the collection and annotation process of the corpus and the extraction manner of the presented lexicon. We enhance the linguistic resources collection by building a sentiment analysis system. We describe in section 4 the followed approach on which the system is based and we evaluate it in section 5. Then, we conclude and give some perspectives.

2. Previous works

A sentiment analysis system has to address the description of both linguistic resources and analysis approaches that form the system basics. Many sentiment analysis systems are based on feature extraction and ignore feature selection algorithms. Thus, many works proceeded to apply feature selection step that allows feature size reducing and improves execution time.

Elnagar *et al*. (2018) introduced a large Arabic review dataset, namely Hotel Arabic Reviews Dataset (HARD). It comprises 490,587 hotel booking reviews crawled from the Booking.com website. The reviews are given along with the corresponding rating score that scales from 1 to 10 stars.

Chetviorkin and Loukachevitch (2014) explored a novel technique to create lexicons. They employed a general lexicon to filter subjective tweets from which they extracted sentimental domain-specific words. They, then, classified polarity based on Markov random field framework. Afterward, they evaluated created lexicon quality which outperformed previous results.

Assiri *et al*. (2018) reported that supervised approaches achieve low performance when dealing with heterogeneous domains as they are domain-dependent. Hence, they proposed a lexicon-based algorithm that weighs words and features domain independence based on a large-scale lexicon for

the Saudi dialect. The weights are assigned to words based on their associations. They also handled some negation and supplication expressions' presence. Moreover, they performed several experiments to evaluate their system performance that reached promising results.

Purpura *et al*. (2019) studied the approaches of emotion classification (e.g. anger, joy, sadness) on a document collection. They found that dictionary-based approaches are the best performing for emotion classification. However, they suffer from limitations such as out of vocabulary terms and cross-domain challenges. In order to overcome these challenges, they employed a supervised approach based on TF-IDF to extract an emotion lexicon and classify diversified-domains' documents. They run experiments on heterogeneous datasets and compared their approach to previous methods.

Rao *et al*. (2015) aimed to determine opinionated parts of speech for given words based on a support vector machine (SVM) in order to classify polarity. They designed a sentiment analysis tool that gathers many feature information.

Saranya *et al*. (2019) considered user rating as a factor that influences customers' decisions when selecting a specific product. They proposed a system that employs a Web application annotating the comments of food recipes based on the corresponding polarities.

Tamilselvi *et al*. (2015) reported the importance of feature selection in improving the performance of a system based on machine learning. They addressed the main challenges of feature selection approaches as well as their fundamental characteristics.

Al Nuaimi *et al*. (2019) suggested preprocessing to enhance real-time decisions that are a challenge when dealing with varied and large datasets. In machine learning, feature selection is an ultimate technique for selecting relevant feature subsets from high dimension features. They presented a methodological introduction of feature selection techniques. They also addressed the strengths and weaknesses of both traditional and current feature selection algorithms.

Tubishat *et al*. (2019) addressed feature selection phase importance for machine learning classifiers when analyzing a huge dataset. They explored Whale Optimization Algorithm (WOA), which is a metaheuristic optimization algorithm. They looked for WOA improvement by using Elite Opposition-Based Learning at the algorithm initialization phase. They also employed an Information Gain filter to select features with WOA using an SVM. They compared the proposed algorithm with six optimization

algorithms as well as two deep learning algorithms. The proposed algorithm does not only reduce feature number but it has an advantage over other algorithms in term of accuracy.

3. Linguistic resources

3.1. The constructed resources

We aim to analyze online data that contain opinionated reviews with targeting both Modern Standard Arabic (MSA) and Moroccan Dialectal Arabic (MDA). We mean by MDA comments those that are expressed by a Moroccan regardless to the region. WhatsApp, Instagram, Twitter, Facebook, and YouTube are considered opinionated data warehouses as they are rich of opinion reviews. Sources vary, so we consult StatCounter[1] in order to get statistics about social media used by Moroccans. Facebook is the most used according to statistics. Hence, we considered it as a data source, and we employed a Facebook crawler[2] to extract comments from Facebook pages. Then, after automatically removing comments written with Latin characters, we manually removed the objective sentences and annotated the extracted comments as positive or negative. The annotation process is not only a time-consuming task but it reduces the collected data size remarkably when filtering opinionated reviews that are written in Arabic language only. During the annotation process, we manually extracted positive and negative terms besides a set of negation words and negation clitics. The process is described in figure1.

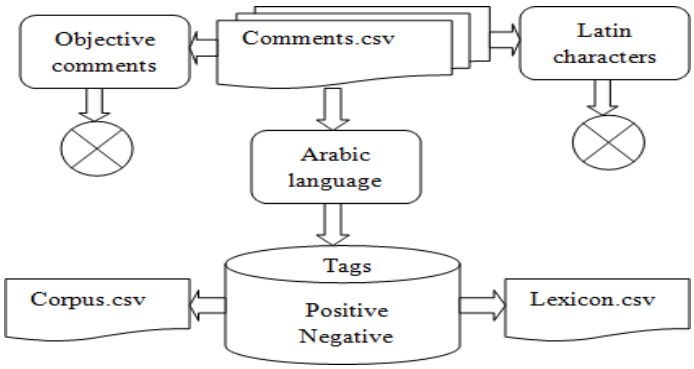


Figure1. Moroccan corpus annotation process

During the annotation process, we considered as MSA comment every comment that all its terms are of MSA type. We employed MDA tag for

[1] http://gs.statcounter.com/social-media-stats/all/morocco

[2] http://scraper.venkatesakumar.com

every comment that contains a set of MDA terms. In Table 1, we give some samples from the constructed corpus.

Sentence_Sentiment	Language	
قادمون ولحماية وطننا الحبيب مستعدون تحت شعار الله الوطن الملك	1	MSA
حرتوا عليهم مزيان وزينوا حرت ههه	-1	MDA
مع الوقت تولي خراب	-1	MDA

Table 1. Samples from the annotated corpus

3.2. Resources preprocessing

The collected corpus was preprocessed by removing all non-Arabic characters, namely, Latin letters, punctuation marks, and digits. The statistics of the corpus are given in Tables 2 and 3.

Tag_Positive	Negative	Total	
Number of tweets	2591	2482	5073

Table 2. Corpus splits by polarity

Corpus	Terms number	Average number of words per tweet
Statistics	62465	12,3

Table 3. Corpus terms statistics

The extracted sentimental terms may occur many times. Hence, we automatically filtered the redundancy between terms. In the following, we give statistics of the obtained lexicons in Table 4 and negation indicators either terms or clitics in Table 5.

Tag_Positive	Negative	Total	
Lexicon terms	2711	3329	6040

Table 4. Lexicon statistics

Tag	Negation terms	Negation proclitics	Negation enclitics
Statistics	167	7	2

Table 5. Negation lexicon statistics

4. Proposed approach

The proposed sentiment analysis system is based on a supervised approach that extracts features from an opinionated dataset, and then filters them based on the following filters:

Term occurrence (TO): it filters lexicon terms by removing those with occurrence lower than five. The threshold five was chosen after performing many tests on the obtained lexicon size. The resulting lexicon contains 493

terms that represent 8.16% of the initial lexicon. We consider this lexicon as initial to our system.

InfoGainAttributeEval (IG): It measures the information gain of an attribute with respect to the class.

CorrelationAttributeEval (CA): It addresses and evaluates the correlation between an attribute and its class.

The input corpus is splited into training and test sets. Then, we extract features from it. We train a model employing an SVM classifier based on the extracted features after selecting pertinent ones and integrating them into a vector that is represented by terms occurrence besides the score of true and false negative and positive terms respectively. Then, we feed characteristic vectors into a classifier to build a model in order to classify unseen data from the test set. The classification process is described in figure2.

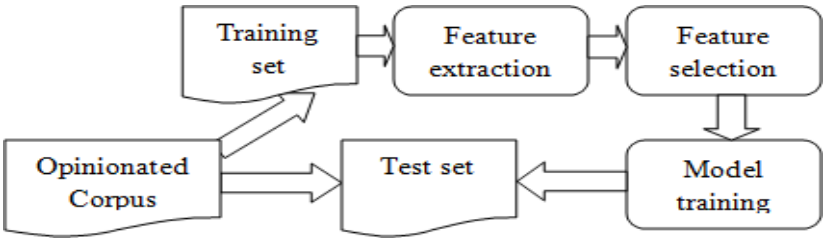


Figure 2. Classification process

5. Experiment results

In this section, we evaluate our system based on the features that were extracted from the crawled and manually annotated corpus. The results are given using percentage split that trains a model on 80% of the data and test it on the remainder as well as 10-folds cross validation (CV), which iterates over all the dataset by altering test folds. Table 6 shows the obtained results based on the occurrence filter using an accuracy metric. % split is sensitive to the training set, thus cross validation may reduce the noise caused by % split test. The obtained results are promising regarding the nature of the collected corpus that implies both modern standard Arabic and Moroccan dialect.

Test option_% split	**CV**	
Accuracy	80.88%	82.65%

Table 6. Occurrence filter accuracy results

In the following, we present the results of the experiments performed on the same corpus by altering filter methods and the number of features to extract (FN). The tests are performed using percentage split test option. The obtained results, in Table 7, either when using IG or CA are very close.

Filter_FN	FN/2	FN/5	
IG	80.98	80.98	80.19
CA	81.08	81.37	80.68

Table 7. IG and CA filters 'accuracy results

From tables 6 and 7, the results show that filtering by occurrence, IG or CA gives concurrent results. Moreover, using IG filter ranks attributes by information gain value, and thus the filtered attributes are those with low information gain value. Whereas, CA filters attributes with a low correlation with the class. When we specify the features' number, both IG and CA diminish not only feature size but the training time as well. From Table 8, there is a significant gain (passing from 100%, when using the initial lexicon without defining the number of features to select, to less than 10%, when using the reduced lexicon) at execution time level when reducing features, which helps to reduce the time of decision-making.

Filter_FN	FN/2	FN/5	
IG	100%	31.08%	9.33%
CA	100%	30.61%	8.20%

Table 8. Execution time according to filters

6. Conclusion and further work

In this paper, we presented both manual and automatic tasks, namely linguistic resources construction and feature extraction and collection processes, in order to build a sentiment analysis system. We point the collection of linguistic resources on both modern standard Arabic and Moroccan dialect. Thus, we filtered comments that don't belong to either of them. Then, the annotation was performed by assigning positive and negative tags to the remainder.

From the collected corpus and during the annotation phase, we extracted both sentimental and negation lexicon words.

The annotated corpus besides lexicons was used to found our sentiment analysis system. Thus, we proceeded to extract features based on these linguistic resources. However, the extracted features are not evaluated according to the carried information. Thus, we followed an approach to

extract the most informative features based on occurrence, information gain and correlation filters.

The employment of IG and CA filters does not only reduce execution time significantly, but it also gives concurrent results to the employment of the initial features filtered by occurrence.

As further work, we want to enlarge the constructed corpus and make a comparison between a set of feature selection algorithms, in order to shed the light on their strengths, weaknesses in terms of vector-size reduction, computation time reduction and their contribution in results readability and interpretation.

References

Al Nuaimi N., Masud M.M., Serhani M.A., Zaki N. (2019). Streaming feature selection algorithms for big data: A survey. Applied Computing and Informatics.

Assiri A., Emam A., Al-Dossari H. (2018). Towards enhancement of a lexicon-based approach for Saudi dialect sentiment analysis. Journal of Information Science, Vol. 44, pp. 184–202.

Chetviorkin I., Loukachevitch N. (2014). Two-Step Model for Sentiment Lexicon Extraction from Twitter Streams. Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 67–72.

Elnagar A., Khalifa Y.S., Einea A. (2018). Hotel Arabic-Reviews Dataset Construction for Sentiment Analysis Applications. In Shaalan K., Hassanien A.E., Tolba F. (Eds.), Intelligent Natural Language Processing: Trends and Applications. Springer International Publishing, Cham, pp. 35-52.

Purpura A., Masiero C., Silvello G., Antonio Susto G. (2019). Supervised Lexicon Extraction for Emotion Classification. Companion Proceedings of the 2019 World Wide Web Conference, ACM Press, San Francisco, USA, pp. 1071-1078.

Rao R., Polepaka S., Rafeeq Md. (2015). Design of Sentiment Analysis System using Polarity Classification Technique. International Journal of Computer Applications, Vol. 125, pp. 22-24.

Saranya K., Roslin R.L., Priyan S.H. (2019). Sentiment Based Comments Rating Approach. International Journal of Innovative Research in Computerand Communication Engineering, n° 2, pp. 1102-1110.

Sutha K., Tamilselvi J.J. (2015). A review of feature selection algorithms for data mining techniques. International Journal on Computer Science and Engineering, Vol. 7, n° 6.

Tubishat M., Abushariah M.A., Idris N., Aljarah I. (2019). Improved whale optimization algorithm for feature selection in Arabic sentiment analysis. Applied Intelligence, 49(5):1688-1707.

DiLiTAL '19, p. 99 - 104

La non voyelle dans la langue arabe

Ismail HADJIR[1], Fatma Zohra BELKREDIM[2], Ahmed ABBACHE[2], Aicha GHOULAM[2]

[1]Faculté des Sciences, Université Dr. Yahia Fares Médéa, Algérie

[2]Faculté des sciences Exactes et d'Informatique, Université Hassiba Benbouali, Chlef, Algérie

hadjir.ismail@univ-medea.dz, {f.belkredim, a. abbache, a.ghoulam}@univ-chlef.dz

Résumé.

Selon la linguistique moderne les voyelles courtes /a, ◌َ/ (fetha) ; / i, ◌ِ/ (kesra) ; /u, ◌ُ/ (dhama) ; /◌ّ /(la gemination, ou chadda) ; /◌ً/ ,/ ◌ٍ/, /◌ٌ/, la nunnation et la non voyelle /◌ْ/ ou soukun sont représentés par des signes diacritiques. Les deux derniers sont appelés les marques de syllabification. Le concept de Sukun existe dans la langue Arabe mais pas dans d'autres langues telles que l'Anglais et le Français. L'importance de ce signe diacritique représenté par un rond, a en langue arabe un rôle phonologique, syntaxique et sémantique. Nous proposons une conception du Sukun basée sur l'expérience x-ray, analogue au zéro que nous connaissons en Mathématiques avec une modélisation de la syllabe.

1. Introduction

L'approche linguistique moderne divise les sons de la langue arabe en consonnes et voyelles. Les consonnes sont au nombre de 29 (Aliane *et al.*, 2011 ; Blachere *et al.*, 1975). Il y a trois voyelles courtes nommées harakat (Wright, 1996) respectivement les sons /a/ (fetha), / i/ (kesra), and /u/ (dhama) qui sont des diacritiques et qui ne font pas partie de l'alphabet de la langue arabe. Il y a aussi trois voyelles longues : /a:/ (alif), / i:/ (ya) , and /u:/ (Waw). Il existe un autre signe diacritique auquel il ne lui ai donné aucun rôle appelé non voyelle (soukun, pause ou stillnes ou absence de voyelle) (Salomon, 2000).

Dans la deuxième partie, on présentera la langue arabe, les consonnes, les voyelles, harakat et Sukun avec comparaison. Dans la troisième partie, une proposition de conception du Sukun basée sur l'expérience x-ray avec une modelisation de la syllabe en langue arabe, puis une conclusion.

2. Langue arabe : consonnes et voyelles

2.1. Consonnes et Voyelles

L'approche linguistique moderne divise les sons de la langue arabe en consonnes (sawamit) et voyelles (sawait) et semi-voyelles (Alkarala, 2004). Les consonnes sont au nombre de 29 (Aliane *et al.*, 2011 ; Blachere *et al*., 1975) ou 28 caractères si on enlève le caractère glottal hamza (Wright, 1996) ou le caractère alif (Ryding, 2005 ; Watson, 2007) des 29 consonnes. Il y a trois voyelles courtes nommées haraket (Wright, 1996) respectivement les sons /a/ (fetha), / i/ (kesra), and /u/ (dhama) qui sont des diacritiques et qui ne font pas partie de la langue arabe. Les signes diacritiques sont des symboles qui sont mis au-dessus et en dessous des consonnes (Ryding, 2005 ; Blachere *et al.,* 1975). En plus il y a trois voyelles longues : /a:/ (alif), / i:/ (ya), and /u:/ (waw), considérées comme des semi-voyelles car elles jouent le rôle de consonnes et de voyelles (Wright, 1996). Il existe d'autres signes diacritiques : la nunnation, la gémination et un autre signe diacritique appelé non voyelle (sukun, pause ou absence de voyelle) (Salomon, 2000) auquel nous nous intéressons particulièrement.

2.2. Haraka et Sukun

Sibawayh dans Alkitaab écrit : ‘‘sukun (repos) n'est pas en mouvement (n'est pas haraka)". Il semblerait que les appellations de voyelles et sukun proviennent de cette définition dans la linguistique moderne. Les haraket sont appelées voyelles courtes et le sukun est appelé non voyelle. De même, il semblerait que l'appellation de non voyelle a conduit à peu de recherche sur ce concept. Néanmoins, pour (Roman, 2011) Sukun est un signe vocalique qui signifie quiescence et le nomme voyelle zéro. AbuZeina *et al*. (2011) appellent Sukun un point silencieux. El harbiyat (2014) et Masmoudi *et al*. (2014) appellent consonne muette si elle possède un sukun. Les linguistes arabes présentent les haraket contre le Sukun. Selon Al- Rummi cité dans (Karaoui *et al*., 2015) sukun est le contraire de haraka.

Pour sibawayh, cité dans (Karaoui *et al*., 2015), la haraka est un outil qui met en mouvement le caractère (harf) qui devient une unité sonore de la parole composée de deux partie : unité stable (consonne) et une unité supplémentaire (la voyelle) qui détermine si la consonne est en mouvement ou pas, d'où la dualité mouvement (haraka) / (non mouvement (Sukun). Il semblerait que Sukun est lié à la consonne, et dans ce cas, la consonne n'est pas en mouvement mais au repos (Sukun en langue arabe a un sens de non mouvement). Autrement dit (harf) est en mouvement (متحرّك) ou au repos (ساكن).

2.3. Comparaison entre haraka et voyelle

L'expérience par X-ray (Karaoui *et al*., 2015) a montré que l'aspect articulatoire du concept haraka partage avec le concept voyelle la même similarité acoustique mais notion de haraka est plus vaste car elle anticipe le processus du mouvement articulatoire. Ainsi haraka dans la langue arabe est un processus d'un mouvement articulatoire et acoustique nécessaire à la fois à la production du son du harf et surtout le passage à la prochaine articulation. Pour le harf sakin conclue qu'il existe un mouvement articulatoire sans production de son. C'est cette expérience qui nous permettra de modéliser le Sukun mais également son rôle dans la langue arabe.

2.4. Rôle du Sukun dans la langue arabe

Dans (Elfakhiri, 2008)[1] : ''Sukun est un morphème important dans la langue arabe. Harakat et Sukun ont la même morphologie et les mêmes fonctions phonémiques liées à la structure morphologique. Ainsi, haraka et Sukun ont les mêmes fonctions (phonémiques, morphologiques et syntaxiques) mais non la même articulation, et le point d'articulation du Sukun n'est pas le même que celui de la haraka. De plus, Sukun est aussi appelé morphème zéro''. Un autre rôle important du Sukun est la location de la position de la consonne dans le conduit vocal. La position de n'importe quelle consonne est localisée en prononçant harf alif suivi du harf C '/ا// C ou harf géminé. En appliquant l'expérience de Karaoui *et al.* (2015), nous considérons le Sukun comme un phonème sans le son(bruit). Par conséquent, le Sukun a une fréquence comme les voyelles /a/, /i/ and /u/ mais sa fréquence est nulle. De plus, Hadj Salah (2005) affirme que le caractère "moutaharik" est toujours considéré plus 'fort', i.e. comme un élément produisant plus d'énergie que harf sakin.

3. Modélisation

En utilisant la conception de Sukun qui est un phonème ayant aucun son acoustique donc ayant une fréquence zéro. Cette conception est analogue à celle du zéro en arithmétique : 'un nombre qui ne représente aucun objet'. De même, nous considérons Sukun comme une entité n'ayant aucun son que nous représentons par /°/. Cette entité peut être appelée « voyelle » sans le son ; comme le zéro un nombre. Dans la pratique, dans l'implémentation,

[1] Né en 1957, Docteur en linguistique à l'université de Tripoli, Lybie depuis 1993

pour diacritiser un texte en langue arabe, nous avons affecté une pondération au Sukun comme pour les voyelles et les autres signes diacritiques (Hadjir *et al*., 2019).

```
Fonction L'emplacement (Haraka est un caractère) : Sortir est un Entier {
Cas (Haraka) {
        ◌ْ      : retourne   2 ;   // Sukun
        ◌َ      : retourne   3 ;   // Fatha
        ◌ُ      : retourne   4 ;   // Dhama
        ◌ِ      : retourne   5 ;   // Kasra
        ◌ً      : retourne   6 ;   // Fathatain
        ◌ٌ      : retourne   7 ;   // Thamatain
        ◌ٍ      : retourne   8 ;   // Kasratain
        ◌ّ      : retourne   9 ;   // Chadaa
  }
  retourne -1;
}
```

Figure 1. Algorithme « Prendre l'emplacement d'un signe diacritique »

3.1. Modélisation de la syllabe en langue arabe

La notion de syllabe est basée sur harf et haraka. 'Harf est le segment minimal de la chaine parlée'. La syllabe de la langue arabe avec les voyelles courtes est CV and CVC (Roman, 2011). La formalisation de la syllabe de la langue arabe CVC est ainsi représentée car le Sukun est considéré comme non voyelle (il n'est pas mentionné ni représenté par un symbole) dans la linguistique moderne.

Avec notre représentation, la syllabe de la langue arabe CVC devient CVC/°/ ce qui rejoint l'idée de Hadj Salah (2005) et Lowenstamn (1996) qui ont démontré que la syllabe de la langue arabe est CV. Nous donnons un exemple :

1. (W1= يَكْتُبُ il écrit)
W1= /y/a//k//t/u/bu/ est modélisé C1/a/+C2/°/+C3/u/+C4/u/ avec
C1=ي, C2=ك, C3=ت, C4=ب avec V1 =/a/ V2=/°/ V3=/u/ V4 =/u/

2. (W2= اكتب écris ! (ordre)
W2= /' //a//k//t//u//b / est modélisé C1 /u/+ C2/°/ C3 /u/ + C4/°/ avec
C1=ا C2=ك C3= ت C4= ب avec V1 =/u/ V2=/°/ V3=/u/ V4 =/ °/

La formalisation actuelle de W1= يَكْتُبُ (il écrit) est CVC-CV-CV ; elle devient CV-CV-CV-CV

La formalisation de W2= اكتب (écris !) est CVC-CVC ; elle devient CV-CV-CV-CV.

4. Conclusion

Dans cette communication, nous avons proposé une conception du phonème Sukun basé sur ses caractéristiques linguistiques et expérience de film radiographique x-ray. Comme conséquence de cette modélisation la structure de la syllabe CVC devient CV-CV ainsi qu'une nouvelle représentation du système acoustique et articulatoire des voyelles qui est représenté par un triangle (se fera en travaux futurs.)

L'utilité de cette recherche peut être étendue à la diacritisation des textes arabes, la segmentation d'un mot, la reconnaissance de la parole et plus généralement dans le traitement automatique du langage naturel.

Références

AbuZeina D., Elshafei M. (2011). Cross Word Modeling for Arabic Speech Recognition. Springer Science & Business Media.

Aliane H., Mazighi A. (2011). Discovering Arabic Structures from texts: what a formal analysis can tell us. International Journal computer Applications in Technology, Vol. 40, n° 4.

Alkarala Z.K. (2004). Vowels in Arabic Language. Edition & Distribution of the Book Elhadith world, Jordan.

Aqil M., Azmi, R., Almadjed, S. (2015). A survey of automatic Arabic diacritization techniques. Natural Language Engineering, Vol. 21, pp. 477-495.

Blachere R., Gaudefroy-Demombynes M. (1975). Classic Arabic Gramma. G. P. Maisonneuve & Larose, Paris.

Elshafei M. (1991). Toward An Arabic Text-To-Speech System. The Arabian Journal for Science and Engeneering, 16(4B):573.

Elfakhiri S. (2008). Journal of linguistics Number 3. Faculty of Art Arabic Language Dept. University of Lybia; Cited in the site http://www.m-a-arabia.com/vb/archive/index.php/t-1384.html

El harbiyat M. (2014). Arabic language Harakat between classical and modern. Amarabek journal. American Arabic academic of the sciences and technology.

Hadj Salah A. (2012). La notion de syllable et la théorie cinético-impulsionnelle des phonéticiens arabes, Al-lisaniyyat, Revue algérienne de linguistique et des sciences et technologies du language-CRSTDLA, n° 17-18.

Hadjir I., Abbache M., Belkredim F. (2019). An Approach for Arabic Diacritization. Natural Language Processing and Information Systems NLDB. Publication du Proceedings Lecture Notes in Computer Science springer.

Karaoui F., Djéradi A., Laprie Y. (2016). The Articulatory Appearance of Haraka in Arabic. Language. 3rd International Conference on Automation, Control, Engineering and Computer Science 4. pp. 643-648.

Lowenstamn J. (1996). CV as the only syllable type and Methods, edited by Jacques Durand.

Masmoudi A., Ellouze K., Estève M., Bougares, Y., Dabbar F., Hadrich Belguith L. (2014). Phonétisation automatique du dialecte tunisien. 30e journée d'étude sur la parole, Lemans France.

Roman A. (2011). Systematic grammar of Arabic Language. L'Harmatan.

Ryding K.C. (2005). A reference Grammar of Modern Arabic Standard. Cambridge University Press.

Solomon I. S. J. (2000). The Beginning of Phonological Terminology in Arabic. Phonoloy: Critical Concepts in Linguistics, Vol. 1. Edited by Charles W. Kreidler.

Watson J. C.E. (2007). The Phonology and Morphology of Arabic. Oxford University Press

Wright W. (1996). A Grammar of Arabic Language. Librairie du Liban, Bey.

DiLiTAL'19, p. 105 - 118

Chunker de la langue arabe basé sur les règles de la grammaire classique

Nabil ABABOU[1], Azzeddine MAZROUI[1], Rachid BELEHBIB[2]

[1]Département d'informatique, Faculté des Sciences, Oujda, Maroc, Université Mohamed Premier

[2]Département de langue et littérature arabe, Faculté des Lettres et des Sciences Humaines, Oujda, Maroc, Université Mohamed Premier

{Nabilaababou, azze.mezroui}@gmail.com, racbel59@hotmail.com

Résumé.

La résolution des ambiguïtés lexicales est un problème complexe, en particulier pour les langues dont la quantité des données annotées disponibles n'est pas assez importante. L'analyse syntaxique partielle ou chunking en anglais consiste à identifier les syntagmes dans le texte et à les étiqueter de manière appropriée. Dans cet article, nous décrivons un Chunker combinant les règles de réécritures et les caractéristiques de la phrase arabe. Le système se fonde sur une structure en trois étapes principales : La création des branches unaires en se basant sur le POSTag Alkalil, la génération des syntagmes possibles, la déduction des combinaisons possibles des syntagmes générés et la sélection de la solution optimale. Nous avons atteint une précision qui dépasse 96%.

1. Introduction

Le Chunking est une tâche de traitement du langage naturel (NLP) qui consiste à diviser un texte en mots corrélées syntaxiquement. Ces syntagmes ne se chevauchent pas, c'est-à-dire qu'un mot ne peut appartenir qu'un seul bloc (Sang et Buchholz, 2000). Il peut être utilisé dans d'autres tâches de traitement plus complexes telles que la traduction automatique, l'extraction d'informations, la synthèse vocale et l'analyse sémantique. Par exemple pour une succession de mot arabe composés seulement des mots, des conjonctions et des prépositions, il suffit seulement de chercher la traduction de chaque syntagme et de la combiner avec les autres afin d'obtenir une traduction de toute la succession comme le montre le tableau 1.

Plusieurs méthodes sont utilisées pour cette tâche. L'étiquetage syntaxique des syntagmes nominaux de Church était l'un de ces premières méthodes, il utilise des fréquences extraites des corpus pour déterminer les limites du syntagme nominal (Church, 1989).

TEXT	**موافقة البنك الدولي على قرض جديد لصالح المغرب بقيمة 033 مليون دولار**			
CHUNKING	بقيمة 033 مليون دولار	لصالح المغرب	على قرض جديد	موافقة البنك الدولي
TRADUCTION DES CHUNKS	D'une valeur de 300 millions de dollars	Pour le Maroc	D'un nouveau prêt	L'Approbation de La banque mondiale
CONCATENATION DES CHUNKS	L'approbation de la Banque mondiale d'un nouveau prêt pour le Maroc d'une valeur de 300 millions de dollars			

Tableau 1. Utilisation du chunking dans la traduction

Abney (1996) a introduit le chunking comme phénomène naturel: "(When I Read) (a sentence), (I Read it) (a chunk) (at atime)."

L'approche d'Abney est devenue le fondement des tâches d'analyse superficielle « shallow parsing ou chunking ». Abney a fait une analyse partielle en utilisant des cascades à états finis, sachant que la cascade à états finis est composé d'une séquence de niveaux. Le syntagme d'un niveau donné est construit à partir d'un syntagme du niveau précédent sans aucune récursivité (Abney, 1996)

Ramshaw et Marcus ont utilisé l'apprentissage basé sur des règles de transformation "TBL : transformation based learning "pour reconnaître les chunks nominaux et aussi les autres chunks du texte pour la langue anglaise (Ramshaw et Marcus, 1999). Ils ont considéré le Chunking comme tâche d'étiquetage en utilisant les étiquettes IOB. Ils ont utilisé B pour désigner le début du syntagme, I pour mentionner le mot à l'intérieur du syntagme et O pour démontrer qu'un mot est à l'exterieur du syntagme. Leurs travaux ont initié une nouvelle idée et des nouvelles recherches sur le chunking sont apparus. Chen (1993) ont proposé un chunker probabiliste basé sur le phénomène introduit par Abney (1996).

Shen et Joshi (2003) ont présenté une nouvelle idée d'étiquetage des données ; au lieu d'utiliser le POS tagging, une nouvelle forme d'étiquetage nommée super tagging (pour exposer plus de dépendances syntaxiques) a été utilisée pour détecter les syntagmes nominaux. Cette méthode d'étiquetage a permis aux chercheurs de gagner environ 1% du Fscore.

Park et Zhang (2003), décrivent une nouvelle approche hybride pour le chunking pour la langue coréenne. Initialement, ils ont implémenté un

chunking basé sur des règles. Ensuite, la technique d'apprentissage basée sur la mémoire MBL est utilisée pour la correction des erreurs. En utilisant uniquement des règles leur système est arrivé à un score de 91,87 par contre l'approche hybride a donné un score de 94,21.

Un Chunks est un seul mot (tête) entouré d'un groupe de mots. Les chunks sont normalement considérés comme un groupe de mots corrélés non récursif. L'arabe étant une langue agglutinante, sa structure morphologique et syntaxique est très complexe. Elle est aussi une langue dont l'ordre des mots est relativement libre mais dans la construction des syntagmes, elle se comporte comme une langue dont l'ordre des mots est fixe. Ainsi, le processus de chunking pour la langue arabe est moins complexe que les autres processus tel que le POS Tagging à condition de prendre en considération les caractéristiques et les lois de la syntaxe arabe.

2. Chunking dans la langue arabe

Nous présenterons dans ce paragraphe quelques chunkers de la langue arabe :

Mohammed et Omar ont développé en 2011 un chunker (shallow parser) de la langue arabe en utilisant une approche basée sur des règles (Mohammed et Omar, 2011). L'analyseur développé par ces chercheurs est basé sur :

- un module d'apprentissage comprenant trois processus : la normalisation, la segmentation et la désambiguïsation ;
- un corpus : les chercheurs ont créé un corpus de 350 phrases arabes dont 280 ont été utilisées pour extraire des règles et les 70 phrases restantes pour tester le modèle ;
- des règles de la grammaire : elles ont été conçues et développées de manière à faciliter l'analyse syntaxique en dépendances. Elles sont définies à l'aide d'expressions régulières et sont au nombre de 150 règles.

Ce système a été testé sur un corpus de 70 phrases et a donné un F-score de l'ordre de 97%.

Fraj et Kessentini ont développé en 2012 un chunker basé sur la décomposition du problème de chunking en deux phases en vue de réduire la complexité et améliorer l'efficacité (Fraj et Kessentini, 2012).

Ils ont supposé que l'utilisation d'une approche d'apprentissage est plus avantageuse que celle basée sur des règles, en avançant que la construction d'une grammaire couvrant toutes les structures syntaxiques de l'arabe (exceptionnelles et spécifiques) est très difficile, voire impossible. Ainsi ils

ont choisi d'envisager une approche modulaire qui repose sur les deux étapes suivantes :

- L'identification des têtes des chunks.
- La reconnaissance des chunks en se basant sur les résultats du POS tagger.

Dans le système Base phrase chunking BPC (Diab, 2009), les chercheurs utilisent comme traits des étiquettes morphosyntaxiques et des n-grammes. Ils ont adopté un schéma d'annotation IOB. Leur approche utilise un modèle de séquence sur SVM. Selon leur article, le composant AMIRA BPC est très rapide et extrêmement précis. Le système donne un F1-score de 96,33%.

Dans l'étude (Khoufi *et al.*, 2014) proposée par Nabil KHOUFI, les chercheurs ont présenté une méthode de chunking des textes arabes basée sur l'apprentissage supervisé. Ils ont utilisé l'algorithme Conditional Random Fields et le Penn Arabic Treebank pour former leur modèle. Pour l'expérimentation, ils ont utilisé plus de 10 100 phrases en tant que données d'apprentissage et 2 524 phrases pour le test. Les expériences ont montré qu''ils ont obtenu des résultats satisfaisants. Ils ont atteint une précision de 96,54%.

Finalement Michael Nawar Ibrahim a proposé un système permettant d'automatiser l'analyse grammaticale des phrases arabes. La tâche de l'analyse de la grammaire arabe a été divisée en trois sous-tâches consistant à déterminer l'étiquette grammaticale, le cas et le signe de chaque unité. Le système proposé a été testé et les expériences ont démontré qu'il permet d'obtenir une précision de 89,74%(Ibrahim, 2015).

3. Description de notre méthode

Le système proposé est divisé en cinq composants principaux : Au début, l'étiqueteur morphosyntaxique dans le contexte attribue une étiquette morphosyntaxique à chaque unité lexicale. Ensuite, le système génère des chunks par des règles de réécritures et enfin le système choisit les bons chunks selon la nature et le nombre des syntagmes dans la phrase arabe. La Figure 1 montre le processus d'exécution de notre algorithme.

3.1. Création des branches unitaires

Dans cette partie le système utilise la sortie de ALkhalil POS (Ababou et Mazroui, 2016) afin de créer des branches unitaires et c'est ici qu'il propose si le syntagme peut être principal ou secondaire (Ababou *et al.*, 2017). Ce syntagme unitaire transfert sa nature à tous les syntagmes qui vont l'accepter par la suite comme tête.

Rappelons qu'un syntagme principal est un syntagme fondamental dans la structure de la phrase, la tête de ce syntagme joue l'une des fonctions syntaxiques déclinées dans le Tableau 2.

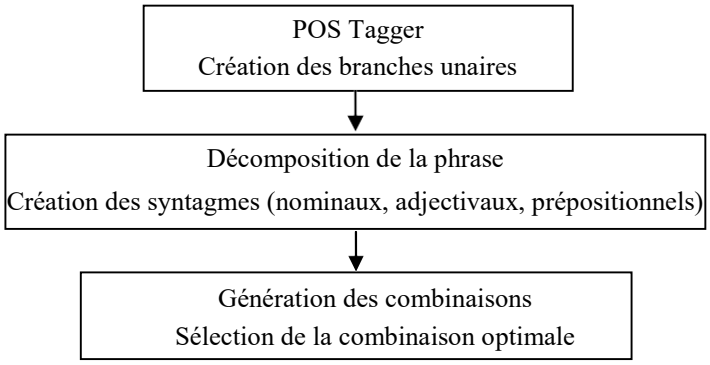

POS Tagger
Création des branches unaires
Décomposition de la phrase
Création des syntagmes (nominaux, adjectivaux, prépositionnels)
Génération des combinaisons
Sélection de la combinaison optimale

Figure 1. Les phases d'analyse

Le prédicat d'une phrase nominale	الخبر
Le sujet d'une phrase nominale	المُبْتَدَ أ
Le sujet	الفاعِل
L'objet direct	المفعُول بهِ

Tableau 2. Syntagmes principaux.

En outre le syntagme secondaire est un syntagme supplémentaire qui enrichit le sens de la phrase, la phrase reste correcte si on l'élimine. La tête de ce syntagme joue l'une des fonctions syntaxique suivante :

Substitut	البدل
Affermissement	التَّوْكِيْد
Adverbe d'état	الحَال
Exception	ألمُسْتَثْنَى
Adverbe du temps ou de lieu	المَفْعُول فِيْهِ
Complément absolu	المَفْعُولُ المُطْلَق
Complément de concomitance	المَفْعُول مَعَه
Complément de cause	المَفْعُول لَه
Spécificatif	التمييز

Tableau 3. Syntagme secondaires

Puisque chaque syntagme unitaire peut devenir la tête d'un syntagme principale ou secondaire et comme nous avons déjà expliqué l'importance du nombre de syntagmes principaux dans la construction de la phrase arabe, Donc nous allons utiliser deux paramètres (P,Q) afin de déterminer le nombre de syntagmes principaux dans chaque phrase.

- Si S(P,Q) 1,0) : le syntagme S est purement principale
- Si S(P,Q) =(1,1) : le syntagme S peut être principale comme il peut être secondaire
- Si C(P,Q) =(5,3) : la combinaison C est composée de cinq syntagmes principaux dont trois peuvent être secondaires.
- [رجل , (P,s)=(1,0)] : tous les syntagmes qui acceptent ce mot comme tête vont nécessairement jouer un rôle principale.
- [عمر, (P,s)=(1,1)] : tous les syntagmes qui acceptent ce mot comme tête auront la possibilité de jouer les deux rôles et c'est dans les phases qui viennent que le système décide la nature de ce syntagme.
- Pour l'étiquette Alam on donne à (P,S) la valeur (1,1) car elle peut jouer le rôle syntaxique secondaire (البدل, Substitut), comme elle peut jouer un rôle principale comme le montre les deux phrase ci-dessous :
- الخليفة عمر عادل
- عمر عادل

3.2. Création des syntagmes

Avant la création des syntagmes (Ababou *et al.*, 2017), le système subdivise la phrase en propositions en utilisant la nature des catégories morphosyntaxiques obtenues par le POSTag. Le tableau 4 comprend les séparateurs de phrase.

SEPARATEUR_TAG	Verbe Nasikh, Particle nasikh, P.PCo, Dz.Ma, conj+Pronom_sujet
SEPARATEUR_ EXPRESSION	لهذه، ونتيجة لهذا، وعلى ذلك، وعلى هذا، فيما يأتي، إذا، لابد، ربما، إنما، وكذلك، ومن هنا، ونتيجة، فحيث، وحيث، حيث

Tableau 4. Les séparateurs entre les propositions

Ensuite, le système crée tous les syntagmes possibles (nominaux, adjectivaux et prépositionnels) pour chaque proposition en se basant sur le phénomène d'annexion (annexion, ألإضَافة) et sur l'ordre des prépositions et des adjectifs

par rapport aux syntagmes nominaux. Le tableau 5 nous montre les règles de réécriture utilisées pour créer ces trois types de syntagmes.

A ⟶ N_No + N_N N_N ⟶ N_No + N_N | N_No	A : Une annexion indéterminée N_No : nom indéterminé
DA ⟶ N_No + D_N D_ ⟶ N_No + D_N | DNo | alam | No.S | NoRP | NoDP | Pronom	DA : Une annexion déterminée DNo: nom déterminé avec l'article (al) No.S: nom déterminé puisqu'il est attaché avec un pronom. Alam : nom propre.
DAdj ⟶ DA + DSifa AdjP ⟶ A + Sifa	Sifa : adjectif indéterminé DSifa : adjectif déterminé AdjP : syntagme adjectival indéterminé DAdjP : syntagme adjectival déterminé
SP ⟶ Prep + AdjP| DAdjP|DA|A	SP : syntagme prépositionnel

Tableau 5. Génération des syntagmes possibles

3.3. Génération des combinaisons

Après la création des syntagmes, le système cherche toutes les combinaisons possibles en se basant sur les paramètres P et S. Il n'accepte que les combinaisons qui respectent le nombre des syntagmes voulus comme le montre le schéma ci-dessous.

Pour une phrase Ph décomposée en k syntagmes (Si)1≤i≤k, nous notons :

$$(C_P, C_Q) = (\sum P_{S_i}, \sum Q_{S_i})$$

CQ n'est autre que le nombre de syntagmes suggérés comme principaux et CP le nombre de syntagmes parmi ces derniers qui peuvent être secondaires.

Exemple :

$(C_P, C_Q) = (5,3)$: la combinaison est composée de cinq syntagmes principaux dont trois peuvent être secondaires.

Nous donnerons par la suite quelques exemples pour illustrer le déroulement de l'exécution.

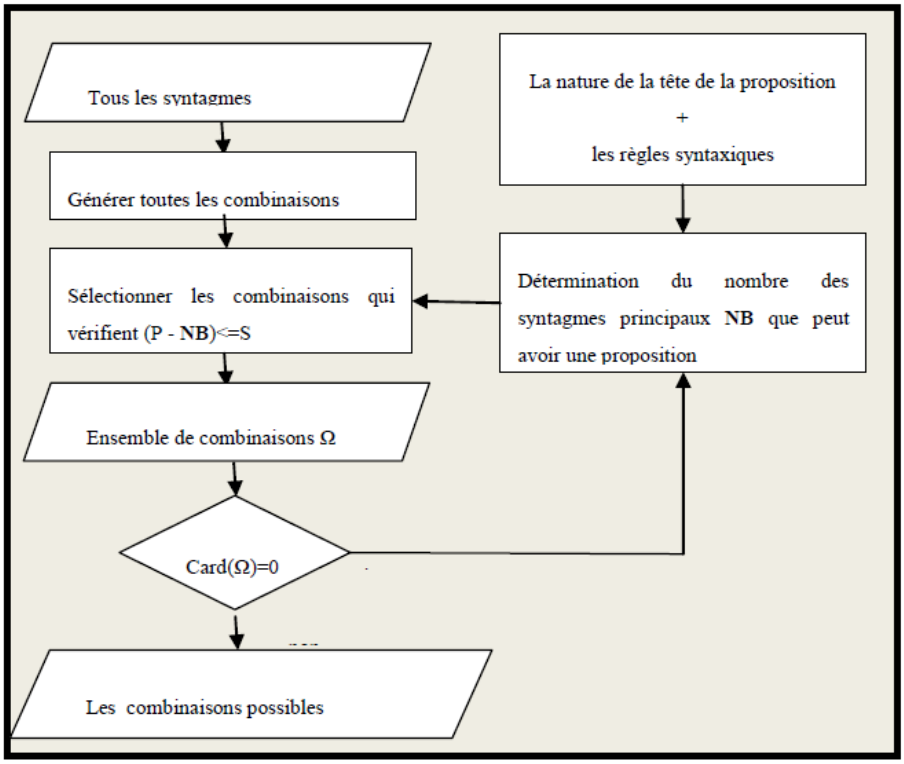

Tous les syntagmes
Générer toutes les combinaisons
Sélectionner les combinaisons qui vérifient (P - NB)<=S
Ensemble de combinaisons Ω
Card(Ω)=0
Les combinaisons possibles
La nature de la tête de la proposition
+
les règles syntaxiques
Détermination du nombre des syntagmes principaux NB que peut avoir une proposition

Figure 2. Génération des combinaisons à partir des syntagmes

Exemple :

" يستعملون في كل أنظمتهم المحلل الصرفي لتوسيم كلمات الجملة" /ystEmlwn fy kl ‘nTmthm AlmHll ALSrfy ltwsym klmAt AljmlP/ (ils utilisent dans la totalité de leurs systèmes l'analyseur morphologique pour étiqueter les mots de la phrase.

En analysant cette phrase, le système propose au début les différentes combinaisons de syntagmes suivantes :

Phrase	Résultats du POS Tagger		Opportunités additionnelles	
	Branche unaire	(P_S,Q_S)	Branche unaire	(P_S,Q_S)
يستعملون	$_0$[VeP.S]$_1$	(0,0)[1]		
في	$_1$[Pr]$_2$	(0,0)		
بداية	$_2$[No]$_3$	(1,0)[2]		
أنظمتهم	$_3$[No.S]$_4$	(1,0)		
لمحلل	$_4$[Sifa]$_5$	(1,1)[3]	$_4$[No]$_5$	(1,0)
الصرفي	$_5$[Sifa]$_6$	(1,1)	$_5$[No]$_6$	(1,0)
لتوسيم	$_6$ [P.No]$_7$	(0,0)		
كلمات	$_7$ [No]$_8$	(1,0)		
الجملة	$_8$ [DNo]$_9$	(1,0)		

Tableau 6. Génération des branches unitaires

En utilisant les règles de réécritures pour générer des syntagmes nominaux, adjectivaux et prépositionnels (voir sections ci-dessus), le système obtient les syntagmes suivants :

Phrase	Syntagmes nominaux			Syntagmes adjectivaux	Syntagmes prépositionnels
يستعملون					
في					في بداية برنامجهم
بداية	بداية برنامجهم				
أنظمتهم					
المحلل				المحلل الصرفي	
الصرفي					
لتوسيم	لتوسيم كلمات		لتوسيم كلمات الجملة		لتوسيم كلمات الجملة
كلمات		كلمات الجملة			
الجملة					

Tableau 7. Génération des syntagmes possibles

[1]Le couple (0,0) est affecté au constituant qui ne jouera jamais le rôle d'un syntagme principal.

[2]Le (1,0) est affecté aux syntagmes principaux.

[3]Le couple (1,1) est affecté au syntagme qui être principale ou être secondaire.

où (i+1) et k désignent dans $_i[\ ]_k$ les positions respectives du premier mot et du dernier mot du syntagme.

Nous présentons dans le tableau suivant toutes les combinaisons possibles :

Phrase	Génération des combinaisons						
يستعملون	يستعملون (0,0)	يستعملون (0,0)	يستعملون (0,0)	يستعملون (0,0)	يستعملون (0,0)		يستعملون (0,0)
في	في (0,0)			في (0,0)	في كل (0,0)		في كل أنظمتهم (0,0)
كل	كل (1,0)	في كل (0,0)	في كل أنظمتهم (0,0)	كل (1,0)			
أنظمتهم	أنظمتهم (1,0)	أنظمتهم (1,0)		أنظمتهم (1,0)	أنظمتهم (1,0)		
المحلل	المحلل (1,0)	المحلل (1,0)	المحلل (1,0)	المحلل الصرفي (1,0)	المحلل الصرفي (1,0)	autres décom-positions	المحلل الصرفي (1,)0
الصرفي	الصرفي (1,0)	الصرفي (1,0)	الصرفي (1,0)				
لتوسيم	لتوسيم (0,0)	لتوسيم (0,0)	لتوسيم (0,0)	لتوسيم (0,0)	لتوسيم (0,0)		لتوسيم كلمات الجملة (0,0)
كلمات	كلمات (1,0)	كلمات (1,0)	كلمات (1,0)	كلمات (1,0)	كلمات (1,0)		
الجملة	الجملة (1,0)	الجملة (1,0)	الجملة (1,0)	الجملة (1,0)	الجملة (1,0)		
(CP,CQ)	(6,0)	(5,0)	(4,0)	(5,0)	(4,0)		(1,0)

Tableau 8. Génération des combinaisons possibles

Pour sélectionner la solution optimale le système utilise plusieurs critères afin de choisir la meilleure solution parmi les propositions générées pendant la phase précédente, on cite à titre d'exemple :

- le système favorise les combinaisons d'un nombre réduits de syntagme
- le système favorise le syntagme (syntagme prépositionnel, جَرٌ وَمَجرورٌ) sur le Substitut (badal,البدل) afin qu'il soit un syntagme principale
- le nombre NB qui vérifie la formule P-Nb <= S est déduit à partir de du type de la proposition et de ses adjacentes

Le système prend en considération les fonctions syntaxiques, le nombre et les positions des syntagmes afin de sélectionner le syntagme secondaire qui sera transformé définitivement en principal.

Dans l'exemple précédent l'étiquette morphosyntaxique de la tête de la phrase « يستعملون » est VeP.S, ainsi la tête est un verbe attaché au sujet. Et puisque « يستعملون » est un verbe transitif, la combinaison optimale doit contenir un seul syntagme principal par conséquence le système doit vérifier la formule P-1 <= S.

Parmi toutes les combinaison générées seulement celles correspondantes à (CP,CQ) = (1,0) qui vérifie la condition P-1 <= S donc cette dernière représente la solution optimale.

4. Évaluation

Nous avons choisi comme corpus de test le livre de Mustapha ben Hamza "الدعوة إلى العامية المسار والأهداف" /AldEwp <lY AlEAmyp AlmsAr wAl>hdAf/ (l'appel à l'usage du dialecte : le parcours et les objectifs) Ce corpus contient 5800 unités lexicales (Ababou *et al.*, 2019). La taille moyenne des phrases du corpus est de l'ordre de 23 unités lexicales et la phrase la plus longue est composée de 73 mots.

Afin d'évaluer les résultats de notre chunker, nous avons employé la mesure Parseval (Black, 1992) utilisée pour juger la validité des constituants. Cette mesure considère un constituant correct si son type et ses frontières sont correctes. Pour mesurer la qualité du chunking, nous avons utilisé la Précision, le Rappel et la F-mesure définis par :

Précision = nombre de constituants corrects sur le nombre total de constituants de l'analyse produite.

Rappel = nombre de constituants corrects sur le nombre total de constituants de l'analyse correcte

F-mesure : combinaison du rappel et de la précision F-mesure = 2. (Précision*rappel) /(précision+rappel).

Nous présentons dans le tableaux 9 les résultats du chunking relatifs aux syntagmes nominaux, adjectivaux et prépositionnels.

	Syntagme nominal	Syntagme adjectival	Syntagme prépositionnel
Précision	0.97680	0.97245	0.9825
Rappel	0.99107	0.96185	0.9706
F-mesure	0.98389	0.96712	0.9765

Tableau 9. Performances du chunking

En analysant ces résultats, nous pouvons avancer les conclusions suivantes :

- Le système réalise de très bonnes performances relatives à la reconnaissance des syntagmes. En effet, la F-mesure pour les trois familles de syntagmes nominaux, adjectivaux et prépositionnels dépasse 96%.

- La faible différence entre les valeurs des précisions et des rappels pour les trois familles reflète les hautes performances obtenues par notre POS Tagger utilisé dans la première phase du système. De plus, en ayant imposé à notre système de générer systématiquement l'étiquette No ou Dno lorsque le POS Tagger génère l'étiquette Sifa ou Dsifa, le système a pu corriger une bonne partie des erreurs dues à l'étiquette Sifa ou Dsifa générée à tort par le POS Tagger au lieu de l'étiquette No ou Dno.

- La distinction entre un syntagme principal et un syntagme secondaire est un avantage considérable, car en exploitant ces informations les combinaisons inutiles qui comportent des faux syntagmes sont rapidement éliminées en faveur des combinaisons qui respectent la condition CP-NB≤CS.

- Les bonnes performances de la phase chunking peuvent aussi être expliquées par le fait que certaines erreurs de niveau inférieur sont corrigées dans le niveau supérieur lors de la composition des syntagmes (Nom et Adjectif). De même, la fonction syntaxique similaire que peuvent jouer les sous-catégories de l'étiquète Nom (Nom, NoRP, NoDP, NomComp) dans les deux types de dépendance (annexion et relation prépositionnelle) contribue également à la précision des syntagmes obtenus.

- En plus des bons résultats obtenus par rapport à ceux de l'état de l'art, notre système utilise des étiquettes et des dépendances propres à la langue arabe et non celles déduites des Treebanks qui utilisent les annotations de l'anglais.

▪ Enfin, une partie importante des erreurs est due à l'ambigüité entre les noms et les verbes trilitères. En effet, le POS Tagger favorise souvent l'étiquète Nom et cela augmente la fréquence d'apparition des syntagmes nominaux dans les résultats du test comparativement à leurs fréquences dans le corpus de test. Ce problème a déjà été soulevé lors de la description des résultats du POS Tagger. Pour les syntagmes prépositionnels, les erreurs sont fortement liées à l'étiquetage de la préposition (tête du syntagme).

5. Conclusion

Nous avons présenté dans cet article un Chunker qui ne nécessite pas de phase d'apprentissage. Il combine les notions des formalismes linguistiques avec les notions et les règles principales introduites par les anciens grammairiens arabes telles que la nature des syntagmes et le nombre de syntagmes principaux dans une phrase. L'utilisation des connaissances de la grammaire arabe est alors apparue primordiale pour permettre une meilleure détection des syntagmes. Les résultats de l'évaluation montrent la performance de notre système. En effet Le système a réalisé des très bons scores, la F-mesure pour les trois syntagmes (nominaux, adjectivaux et prépositionnels) a dépassé 96%.

Références

Ababou N., Mazroui A. (2016). A hybrid Arabic POS tagging for simple and compound morphosyntactic tags. International Journal of Speech Technology, 19(2):289-302.

Ababou N., Mazroui A., Belahbib, R. (2019). Elaboration of a treebank for Arabic language. International Conference on Advanced Technology for Humanitarian Sciences, Presented at the FARAH HOTEL, Rabat, Morocco. FARAH HOTEL, Rabat, Morocco.

Ababou N., Mazroui, A., Belehbib, R. (2017). Parsing Arabic Nominal Sentences Using Context Free Grammar and Fundamental Rules of Classical Grammar. International Journal of Intelligent Systems and Applications, Vol. 9, n° 8.

Abney S. (1996). Partial parsing via finite-state cascades. Natural Language Engineering, 2(4):337-344.

Black E. (1992). Meeting of interest group on evaluation of broad-coverage grammars of English. LINGUIST List 3.587.

Chen K., Chen H. H. (1993). A probabilistic chunker. Proceedings of Rocling VI Computational Linguistics Conference VI, pp. 99-117.

Church, K. W. (1989). A stochastic parts program and noun phrase parser for unrestricted text. International Conference on Acoustics, Speech, and Signal Processing, pp. 695-698. IEEE.

Diab, M. (2009). Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking. 2nd International Conference on Arabic Language Resources and Tools.

Fraj F. B., Kessentini M. (2012). Combinatorial classification for chunking Arabic Texts. International Journal of Artificial Intelligence and Applications, Vol. 3, n°5.

Ibrahim M. N. (2015). Statistical Arabic grammar analyzer. International Conference on Intelligent Text Processing and Computational Linguistics, pp. 187-200. Springer.

Khoufi N., Aloulou C., Belguith L. H. (2014). Chunking Arabic texts using conditional random fields. IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA), pp. 428-432.

Mohammed M. A., Omar N. (2011). Rule based shallow parser for Arabic language.
Journal of Computer Science, 7(10):1505-1514.

Park S.-B., Zhang B.-T. (2003). Text chunking by combining hand-crafted rules and memory-based learning. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Vol. 1, pp. 497-504. Association for Computational Linguistics.

Ramshaw L. A., Marcus M. P. (1999). Text chunking using transformation based learning. In Natural language processing using very large corpora, pp. 157-176, Springer.

Sang E. F., Buchholz S. (2000). Introduction to the CoNLL-2000 shared task: Chunking. arXiv Preprint cs/0009008.

Shen L., Joshi, A. K. (2003). A SNoW based supertagger with application to NP chunking. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Vol. 1, pp. 505-512. Association for Computational Linguistics.

DiLiTAL'19, p. 119 - 124

IRCAM-Tifinagh handwritten character recognition using deep learning

Dadi EL WARDANI

SDIC team, Laboratory of Applied Sciences, National School of Applied Sciences of Al-Hoceima, University of Abdelmalek Essaadi

w.dadi@uae.ac.ma

Absract.

In this paper, we exploit the benefits of the deep learning approach to design an efficient system of Amazigh handwritten recognition. Indeed, this approach has proved a greater efficiency in the various domains, especially recognition tasks. However, to take full advantage of this approach it's necessary to construct an adequate dataset of training and testing that represent faithfully the concerned problem. To this end, we have prepared our dataset of 102 writers each one contains 33 characters of IRCAM Tifinagh. Inspired by the MNIST database, the set of characters is size normalized and centered in a fixed-size image. The r*esulting is a* grey level image of size 28x28, where the black color is the non-color of the character. The number of images produced after this preprocessing step is 3,366.

1. Introduction

Handwriting recognition is an active research area in pattern recognition and computer vision. This domain has several applications fields such as documents digitizing and archiving, postal address interpretation, signature verification.... Currently, many Smartphone applications are based on handwritten as input by touching the screen through a stylus or finger. These applications are used for example to take notes.

In recent years, this domain has received much attention in academic areas and it has become one of the most popular research subjects because of developing the new classification and recognition approach. Indeed, the first artificial pattern recognizers system to achieve human competitive performance was performed by Yann LeCun (LeCun *et al*., 1998) on the famous MNIST handwritten digits' dataset. This system is well known in the literature by Convolutional Neural Network(CNN) which is a kind of deep

learning approach. This last has proven very interests and effective results in classification and retrieval tasks.

In this paper, we are interested in Tifinagh character which is the Amazigh language alphabet. Indeed, the Amazigh language is widely used in North Africa in countries such as Morocco, Algeria, Libya, Tunisia, Mauritania, etc. This language and its alphabet Tifinagh have been recently recognized. The Tifinagh code approval by the International Organization for Standardization (ISO) that was in 2004 has allowed its use in several computer systems nationally and internationally. The Tifinagh writing is currently integrated by major companies specializing in the production of software, which will facilitate its use especially in terms of Internet and text processing.

The emergence of Tifinagh character in the digital world has created an increased need for several specific computer systems like those developed for other languages, namely translation systems, automatic document filing systems, security systems, automatic processing of administrative documents, postal automation. These systems can use a handwritten character recognition system.

To promote machine learning and pattern recognition research, several standard databases have emerged in which the handwritten digits are preprocessed, including segmentation and normalization, so that researchers can compare recognition results of their techniques on a common basis. This dataset needs to represent a sufficiently challenging problem to make it both useful and to ensure its longevity. For Tifinagh handwritten dataset, at our knowledge two databases are proposed (Bencharef, 2015; Es Saady, 2011). The two datasets are developed basing on 33 characters of IRCAM-Tifinagh adopted by the Royal Institute of Amazigh Culture, Morocco. These two databases contain an insufficient number of examples and also the examples that do not represent every possible case. Furthermore, the existence of a varied suite of benchmark tasks is important in allowing a more holistic approach to assessing and characterizing the performance of an algorithm or system.

In this context, we propose a new dataset of Tifinagh handwritten. To construct this dataset, we have selected a normal sample of 102 writers that represent different age and education levels. To outperform a simple-to-use dataset, our preprocessing operation is inspired by the famous MNIST dataset (LeCun *et al*., 2007), where the set of characters is size-normalized and centered in a fixed-size image. The resulting is a grey level image of size 28x28, where the black color is the non-color of the character. This is a relatively simple to-use dataset for people who want to try machine learning

techniques and pattern recognition methods, especially the deep learning approach, on real-world data while spending minimal efforts on preprocessing and formatting.

The proposed dataset is used to develop an efficient recognition system basing on deep learning. Indeed, this approach and in particular the convolutional neural networks (CNNs), have proven to be very efficient in the most recognition task.

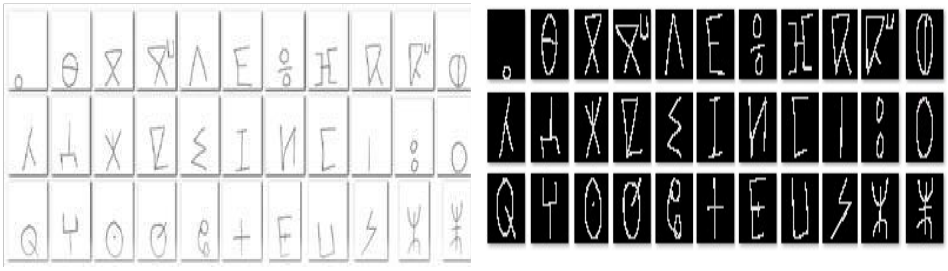
2. Dataset

In this paper, we propose a new pre-processed dataset of Tifinagh handwritten character. Our proposed database is inspired by the famous MNIST dataset, where the set of characters is size-normalized and centered in a fixed-size image. The resulting is a grey level image of size 28x28, where the black color is the non-color of the character. This is a relatively simple-to-use dataset for people who want to try machine learning techniques and pattern recognition methods, especially the deep learning approach, on real-world data while spending minimal efforts on preprocessing and formatting.

To preprocessing this data, we have defined a particular and general treatment. The particular one concern especially the first alphabet which is a = " ⴰ". Indeed, in Tifinagh writing, this character takes a smaller size than the other alphabets especially on the vertical level (cf. Figure 1).

Figure 2. The 33 IRCAM-Tifinagh alphabets

The different pretreatments applied are framing, adjustment and standardization. The main objective is to produce images of the same size (28 * 28), taking inspiration from the MNIST database.



(a) (b)

Figure 3. IRCAM-Tifinagh handwritten sample: (a) before and (b) after preprocessing

Our preprocessing steps are performed as follow:

As the characters do not fill the entire image, the region around the actual character is extracted.

- Then it is placed and centered into a square image with the aspect ratio preserved.
- The region of interest is padded with a 2 pixels border when placed into the square image.
- Finally, the image is down-sampled to 28 × 28 pixels using bi-cubic interpolation.

3. Training and test using CNN

For the CNN architecture, we have used a simple one constructed by two convolutions layers and 2 layers of max-pooling.

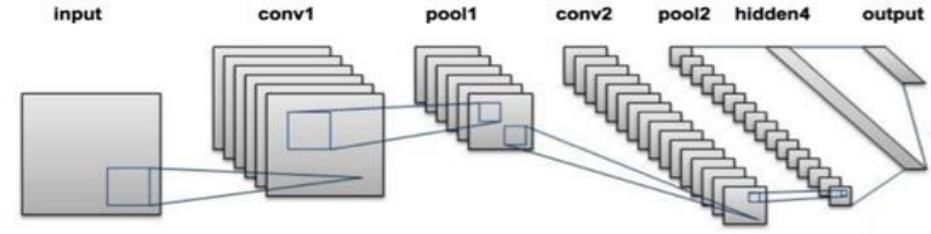


Figure 4. The used CNN architecture.

To perform tests, we have split our dataset as follows: 86% (88 samples) for the train split and 14% for the test split.

The training process is performed on a PC equipped with NVDIA GeForce GT 740. Table 1 and Figure 4 show the accuracy obtained after 100 epochs

of training and testing, which is a very important and interesting precision. Indeed, the obtained results for both operations are very close to 100%. In addition, as shown in Figure 4, with more eras, the loss and accuracy of the model on the training and test data converged, thus making the model stable.

	Top-1	Top-5
Train	0.9969	1.0000
Test	0.9475	0.9960

Table 1. Accuracy

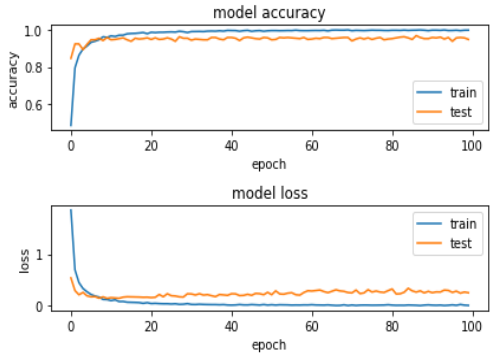


Figure 5. Evolution of accuracy and loss for 100 epochs

4. Conclusion

In this paper we have presented a dataset of IRCAM-Tifinagh handwritten character constructed of 102 samples. This dataset is used to develop a recognition system of the concerned alphabet using deep learning approach. Our recognition system presents an important accuracy.

The number of samples presented in this work still insufficient, so it's crucial to construct a large database. To this end, our solution consists to use the data augmentation. This technique serves not only to have a large database but also to have a system which characterized by the property called invariance. More specifically, a CNN can be invariant to translation, viewpoint, size or illumination (Or a combination of the above). Moreover, and for more challenging dataset, in the next work we want to add some other writers as perturbed sample. This perturbation consists of adding characters with missing parts.

References

Bencharef O, Chihab Y, Mousaid N, Oujaoura N. (2015). Data set for Tifinagh handwriting character recognition. Data in Brief. Vol. 4, pp. 11-13.

Cohen, Saeed Afshar, Jonathan Tapson and Andre van Schaik. (2017). EMNIST: an extension of MNIST to handwritten letters. https://arxiv.org/abs/1702.05373.

Es Saady Y, Rachidi A, El Yassa M, Mammass D. (2011). AMHCD: A Database for Amazigh Handwritten Character Recognition Research. International Journal of Computer Applications. 27(4): 44-48.

LeCun Y., Corinna C., Burges J.C. (2007). MNIST handwritten digit database, http://yann.lecun.com/exdb/mnist/

LeCun Y., Bottou L., Bengio Y., Haffner P. (1998). Gradient-Based Learning Applied to Document Recognition, Proceedings of the IEEE, 86(11):2278-2324.

DiLiTAL'19, p. 125 - 131

Using a novel approach for automatic concepts extraction from Arabic texts

Aicha GHOULAM[1], Ahmed ABBACHE[1], Fatma Zohra BELKREDIM[1], Ismail Hadjir[2]

[1]Faculté des sciences Exactes et d'Informatique, Université Hassiba Benbouali, Chlef, Algérie

[2]Faculté des Sciences, Université Dr. Yahia Fares Médéa, Algérie

{a.ghoulam, a.abbache, f.belkredim}@univ-chlef.dz, hadjir.ismail@univ-medea.dz

Abstract.

Arabic language is one of enriching Semitic languages that has complex morphological, grammatical, and semantic aspects. Due to this complexity, automatic Arabic concept extraction is difficult. Automatic concepts extraction methods analyze a document to determine the important domain terms which can be single word or multi-word term. In this paper, we present an approach to extract automatically concepts from Arabic documents. This approach combines statistical analysis and Arabic WordNet (AWN) information.

1. Introduction

The concepts extraction is a process to identify a list of terms that capture the most important ideas or topics covered in a document. Crangle *et al*. (2004) defined concept extraction as follows: "concept extraction is the process of deriving terms from natural language text that are considered representative of what the text is about. The terms are natural language words and phrases which may or may not themselves appear in the original text".

Concepts extraction has been used in many of natural language processing (NLP) applications, such as relation extraction, ontology construction, information retrieval systems, digital library searching, document clustering, and text summarization (Rose *et al*., 2010). Although concepts extraction is very important for a large of applications. Therefore, developing tools to

automatically extract concepts from Arabic texts are increasingly needed mostly with the enormous amount of Arabic documents written in natural language and available in electronic forms.

Manual concept extraction is time consuming process and it is not practical. The automatic concept extraction methods often analyze a documents to determine the important domain terms, which can be a single term or multi-word term.

In this paper, we developed a system that automatically extract concepts from Arabic texts. This system based on a method that combines document's statistics and the used of external resource which is Arabic WordNet (AWN). In the document's statistics, all the N-grams that may be considered as concepts are extracted from the documents. We calculate then the frequency of each N-gram and the ones that appear more frequently and exist in AWN are then selected as concepts.

This paper is organized as follows. In section 2, we present related works to Arabic concept extraction. Section 3 describes the AWN resource. Section 4 describes the proposed system and our contribution to extract automatically concepts from Arabic texts. Finally, this paper ends with conclusions and some perspectives for future works.

2. Related works

Among the related works that exist in the Arabic concepts extraction field we show and explain some recent works in this sense:

In the work given by (Khaled *et al*., 2018) to understand the word meaning and solve the disambiguation sense problem, they proposed method to extract the Synset of concepts and word to make their NLP system able to understand the semantic meaning of the text. They used the AWN for the identification of the semantic relationship using concept Synset extraction.

Al-Arfaj and Al Salman (2016) give a summary of the Arabic term extraction methods from Arabic texts. They cited some works and approach from 2008 to 2014. The methods they cited are categorized according to the type of extraction and filtering technique.

Bouhriz *et al*. (2015) proposed an approach to represent the semantic content of Arabic texts, they extract Arabic concepts from text using AWN. the extracted concepts are submitted to a formalism of analysis of formal concepts to obtain reduced and relevant concepts describing the semantic content of the text.

(Najadat and Rawashdeh, 2015) proposed an Automatic Concept Extractor from Arabic corpus, in the first step they identified nouns and in the second step they identified the nouns that represent concepts out of the nouns, which represent anything other than concepts such as names and properties.

Unsupervised approach for automatic keyword extraction from Arabic documents is proposed in (Awajan, 2014), his approach combines statistical analysis and linguistic information, the statistical analysis based on the detection of all N-grams and the linguistic information is to replace the word of the N-grams with their roots for the derived words and with their stems for the non-derivative words.

3. Arabic WordNet

Arabic WordNet (AWN) (Elkateb *et al*., 2006; Black *et al*., 2006) is a large lexical database of Arabic concepts, words, and the semantic relationships among them. AWN is one of the resources for the general Arabic language available online. The AWN could be defined as a set of Arabic words that linked with each other through semantic relationships among these words. The AWN is organized as a structure of Synsets (or concepts), which means a sets of synonyms and pointers describing relationships to other Synsets and each word can belong to one or more Synsets.

In general WordNet considers the most important for several applications of information retrieval and NLP. Recently, the researchers are still updating and development the WordNet to be Multi-language WordNet to become a Global WordNet (Khaled *et al*., 2018).

4. System architecture and proposed approach

Figure 1 shows the global architecture of the system. the system proposed consist to extract automatically the concepts from Arabic texts. Our system has two major phases: Pre-processing phase, and extraction of concepts phase (application of AWN resource and N-grams weighting).

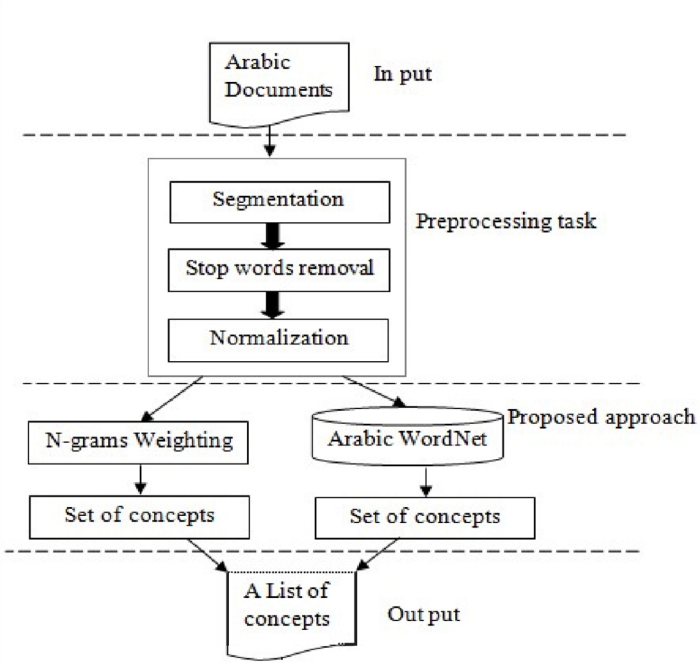


Figure 6. Architecture of the proposed system

4.1. Preprocessing phase

This phase is an essential process; it has several subtasks as follow:

4.1.1. Segmentation

It is necessary to properly delimit the Arabic documents into meaningful units. Most natural language processing solutions expect their input to be segmented into sentences and each sentence into tokens.

4.1.2. Stop words removal

This process is to remove the common Arabic words that considers as insignificant words or that does not carry significant meaning. The result of this subtask is reducing text. It is a list of all the words of liaison and articulations such as: pronouns (الضمائر المنفصلة), prepositions (حروف العطف), conjunctions (حروف الجر). For example: في، على، من، إلى، عن، هذا، هذه، بين، بعد، انه، التى، أن، مع، الذي، لم، ما، منذ،...

4.1.3. Normalization

This process is applying to the reduce texts. It represents a task to transfer the inconsistence text to be more consistency. The normalization task was used in Arabic language to remove the diacritics marks, numbers and normalize the other specific characters. The result of this subtask is a set of normalized tokens. we apply to the normalized text statistical analysis; we calculate N-grams weighting and AWN resource.

4.2. The proposed approach

The proposed approach combines statistical analysis and AWN resource information. In the statistical analysis we extracted all N-grams that exist in the normalized text and then we calculate the weight of each N-grams. We look for the concept that corresponds to each term in the AWN.

4.2.1. The statistical analysis

The normalized text is processed to extract all the possible N-grams that constitute a sequence of adjacent words with a maximum length of three words (N=3). By analyzing the evaluation Arabic corpus, we felt that most of the concepts in this corpus we took do not make sense when N> 3, which is why we want to choose N = 3 as a maximum length of three words.

In this subtask, we calculate the weight of all the possible N-grams and the ones that appear more frequently are then selected as concepts. To calculate the weighting, the following methods are used:

Term frequency (TF) measure: is the percentage of occurrence of the term in a document.

$$TF = \frac{\text{Number of occurrences of a term in a document}}{\text{The sum of the occurrences of all the terms that appear in the document}}$$

Term Frequency-Inverse Document Frequency (TF*IDF): is a weighting method often used in information retrieval. the calculation of TF*IDF can only be done if we have introduced a set of documents. The formula is as follows:

$$TF * IDF = TF * Log = \frac{\text{total number of document}}{\text{number of document containing the term}}$$

The result of this subtask is a set of concepts, the N-grams weighting is considered as concepts by defined frequency threshold (>=3).

4.2.2. Application of AWN resource

We apply the AWN resource to look for the existing of the concept that corresponds to the tokens of the normalized text. If it exists in AWN, it considered as a concept else is it not a concept.

4.2.3. The combination approach

The approach that we proposed combine two phases; the statistical analysis and the application of AWN. The result of the first phase is a set of concepts that appear more frequently in the Arabic document, and in the second phase of application of AWN we considered all tokens that exist in AWN as a concept. For the combination approach we have taken the union of the two phases, so we supposed that all tokens that more frequently appear in the Arabic documents and appear in AWN are considered as concepts.

5. Conclusion

In order to represent the semantic content of the Arabic texts, it is necessary to adopt approaches to extract automatically concepts which is considered as a way of representing this content. The approach proposed in this paper is to combine approach that exploits AWN, and statistical analysis. We plan as future work to experiment the approach that we proposed by using Arabic corpora and standard metrics of recall, precision and f-score to evaluate our concept extraction system.

References

Al-Arfaj A., Al Salman A. (2016). Towards Concept Extraction for Ontologies on Arabic language. International Journal on Islamic Applications in Computer Science and Technology, 4 (4): 9-19.

Awajan A. (2014). Unsupervised Approach for Automatic Keyword Extraction from Arabic Documents. The 2014 Conference on Computational Linguistics and Speech Processing.

Black W., Elkateb S., Rodriguez H., Alkhalifa M., Vossen P., Pease A., Fellbaum C. (2006). Introducing the Arabic WordNet Project. http://www.globalwordnet.org/AWN/meetings/GWApaper.pdf

Bouhriz N., Benabbou F., Benlahmer H. (2015). Une approche d'extraction basée sur la sémantique contenue dans un texte arabe. La troisième journée sur les Technologies d'Information et de Modélisation.

Elkateb S., Black W., Vossen P., Farwell A., Fellbaum C. (2006). Arabic WordNet and the Challenges of Arabic. http://www.mt-archive.info/BCS-2006Elkateb.pdf

Crangle C., Zbyslaw A., Cherry M., Hong, E. L. (2004). Concept Extraction and Synonymy Management for Biomedical Information Retrieval. In Proceedings of the Thirteenth Text REtrievel Conference.

Khaled M., Al-Maghasbeh A., Pouzi bin Hamzah M. (2018). An Automatic Approach of Concepts-Synset extraction from Arabic Wordnet. Global Journal of Human Social Science XVIII(V):45-49.

Najadat H., Rawashdeh A. (2015). Concept Extraction from Arabic Text Based On Semantic Analysis. Proceedings of the The International Technology Management Conference, Antalya, Turkey. ISBN: 978-1941968-11-6 ©2015 SDIWC.

Rose S., Engel D., Cramer N., Cowley W. (2010). Automatic keyword extraction from individual documents. Text Mining: Applications and Theory edited by Michael W. Berry and Jacob Kogan, John Wiley & Sons, Ltd. pp. 3-20.

DiLiTAL'19, p. 133 - 141

Injecting linguistic features into Arabic-English factored translation model

Safae BERRICHI, Azzeddine MAZROUI

Department of computer Science, Faculty of Sciences, Mohamed First University, Oujda, Morocco

{berrichi.safae, azze.mazroui}@gmail.com

Abstract.

Phrase-based statistical machine translation (PBSMT) has overcome several challenges in word-based translation, which allowed a significant improvement in the final translation quality. However, these improvements are still restricted in terms of translation between an agglutinative and highly inflected language and another that which lacks these features, as in the case of Arabic and English languages. In addition, these limits are highlighted for poorly endowed languages. We have developed, in this paper, factored models for Arabic/English translation, which are an extension of the PBSMT models. These models allow integrating linguistic information (lemma, PoS tag...) of the source language in order to improve the translation performance. The various proposed factor-based models were compared with the standard model of the PBSMT system. The results obtained in terms of BLEU score show that the use of linguistic features leads to a significant increase in translation over the standard model.

1. Introduction

Machine Translation (MT) has been the subject of several studies in recent decades. Deep Learning, which represents the most dominant approach to machine translation nowadays, is based on neural networks and requires a high quantity of training data (Koehn and Knowles, 2017). However, the statistical approach continues to be widely used for poorly endowed languages such as Arabic (Koehn *et al.*, 2003). It leverages artificial intelligence knowledge and probabilistic computations. This method is trained on parallel corpora of sentences from the source language with their translations into the target language. During the training phase, the system generates probabilistic models that subsequently serve to identify for a given

sentence of the source language, its translation into the target language with a maximum likelihood score.

Since the Statistical Machine Translation (SMT) is only based on probabilistic computations, its performance is frequently restricted when two languages are not from the same language family. To improve these performances, an external pre-processing step is necessary. It involves the incorporation of additional linguistic information into the training data, which may be morphological, syntactic or semantic. For example, incorporating lemma information (the canonical form of the word) into the translation model can overcome the problem of missing or scarcity of some words in the training corpus.

In this study, we used several items of linguistic information from Arabic words to generate factored models that improve the translation quality between Arabic and English. Some of the information employed includes the surface form (word), its lemma (which considers all the inflected forms of a word as a single unit), and the PoS tag that helps to disambiguate the words. We examine the results of each model individually as well as a combined model.

The present document is organized as follows. Section 2 reviews previous work on MT to/from Arabic and factored translation systems. Section 3 discusses the Factored Machine Translation approach. Section 4 outlines the corpora used in the training and testing phases, and describes the tools used in the translation models. Section 5 reports the evaluation results, and section 6 presents a conclusion and some perspectives.

2. State of the art

This section reviews some related work that has incorporated linguistic knowledge into translation systems. (Bhattacharyya, 2017) studied the impact of enriching the source language with different morphosyntactic information (morphological tags, PoS tags, syntactic tree, etc.) of words to improve the translation performance from English to Hindi and Marathi. The results obtained have shown that the incorporation of this information reduces the number of Out Of Vocabulary words (OOV words) and improves the fluency of the translation. (Aqlan *et al.*, 2019) also investigated the benefit of injecting linguistic features into the source language using factored models, and comparing them with the standard PBSMT model. The evaluation was carried out on different characteristics for the Arabic language (lemma, PoS tag ...). The first treatment used consists of using different tokenization schemes for the preprocessing of the Arabic language with a simple tokenization for the Chinese language as the target language.

Then, the use of linguistic factors to annotate the corpus was applied as a second treatment.

The results obtained illustrate that factored systems have improved translation performance in terms of the BLEU score compared to standard systems. In addition, the factored model (PoS) achieved a better score than the other two factored models (lemma) and (lemma+PoS).

For factored translation models between Arabic and English, the research works are limited. In (Badr *et al.*, 2008), the authors used additional features to build a factored model for the English-Arabic direction. According to the segmentation of Arabic words, the system applied a factored model that annotates the English language with PoS tags and the Arabic language with stem and PoS tags. In the opposite direction Arabic-English, Rajah *et al.* (2016) have developed an approach that injects PoS tags and Combinatory Category Grammar (CCG) into the English language to form an integrated model (PoS+CCG). The results showed that the integration of these factors on the different grams of the language model (LM3, LM5, LM7 and LM9) led to a significant improvement in translation quality compared to the standard PBSMT system. Other previous studies have examined the impact of linguistic factors on the new generation of machine translation, such as (Sennrich and Haddow 2016).

In contrast to this previous work, our approach allows us to test the injection of morphosyntactic information (lemma, PoS tag) for Arabic-English translation as an approach not previously applied in previous studies. The preliminary results obtained were carried out in the PBSMT system, which remains competitive with Neuronal Machine Translation (NMT) due to the scarcity of data for the pair (Arabic, English) (Koehn and Knowles, 2017).

3. Factored translation model

The PBSMT system is restricted to the correspondence between phrases and do not include linguistic knowledge of languages. The phrase-based translation model generated by this system is based on noisy channel models. The formula defined in equation (1) is used to detect the target sentence Tbest having a maximum probability among the sentences T that can translate a given source sentence S:

$$(1) \qquad T_{best} = \underset{T}{\operatorname{argmax}}\, P(T/S) = \underset{T}{\operatorname{argmax}}\, P(S/T)P(T)$$

This formula consists of two probabilistic models. P(T) is a probability relative to the target language estimated by the n-gram language model, and *P(S/T)* is a probability estimated by the translation model through splitting the sentences S and T into phrases. The approach used in this work is based

on the factored model as an extension of the PBSMT (Koehn and Hoang, 2007). This model makes it possible to incorporate linguistic information into the source language and/or the target language before producing the translation and language models. In this paper, we focus on annotating the source language corpus (Arabic). We have incorporated some features like lemma and PoS tag so that surface forms that were tokens become a vector of factors providing linguistic knowledge. The difference between the standard and factored system representations is illustrated in Figure 1.

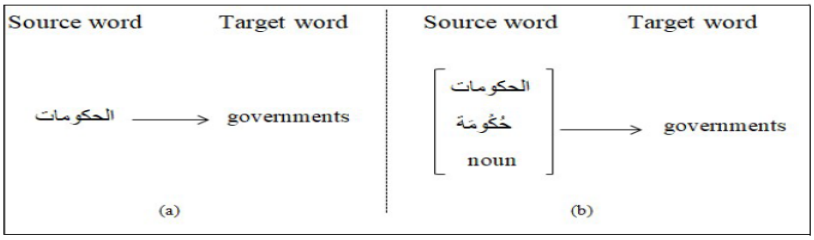


Figure 1. Example of word annotation in translation models. (a) Standard model (b) Factored model

After pre-processing on the training corpus, which annotates the source language by different factors, an alignment step is applied to generate the alignments between the words. These alignments are, then, used to extract and estimate the translation probabilities of the phrases (translation model). Finally, a decoding mechanism will be used to combine the translation model and the language model, and consequently produce the maximum translation of a new input source sentence. Figure 2 outlines our factored model.

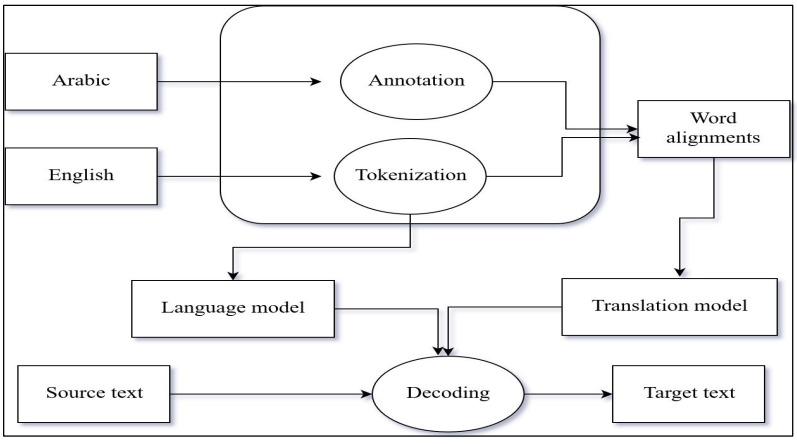


Figure 2. Architecture of our factored model

4. Corpora and tools used

Our focus in this section is the evaluation of the different translation systems. We start by introducing the collected data and the used tools, and then we define the various configuration parameters of the translation system.

4.1. Input data

In our experiment, we collected English-Arabic parallel corpus from the United Nations (UN) documents (Ziemski *et al.*, 2016). To build the translation and language models, we randomly extracted 66585 parallel sentences representing 97% from this corpus. A cleaning script was applied in the training corpus to eliminate all sentences with sizes greater than 100 words. The new training corpus contains 63355 sentences. For the rest of the corpus, we randomly decomposed it into a development corpus composed of 1000 and a test corpus composed of 1400.

Table 1 summarizes the statistics for these three corpora.

Corpora_Number of sentences		Number of words	Vocabulary size	
Training	English	65 355	1 655 691	28 987
	Arabic		1 486 171	32 215
Development	English	1000	30 541	4506
	Arabic		27 191	7757
Test	English	1400	42 135	5366
	Arabic		37 588	9435

Table 1. Statistic details of the three corpora

4.2. Tools

4.2.2. Linguistic features

The annotation of the source language corpus (Arabic) by linguistic features was performed by the MADAMIRA system (Pasha *et al.*, 2014). This system analyzes the word in its context by applying SVM and language models to predict different morphological characteristics of the Arabic word. The characteristics used in our experiments are the lemma and the PoS tag.

For the English language, which is the target language, we have limited to a simple tokenization using the Stanford tool (Manning *et al.*, 2014).

4.2.3. Translation system

All the translation models used in this work were built by the statistical translation system "Moses" (Koehn *et al.*, 2007), which is an open source MT toolkit. For word alignments, we used the GIZA+++ toolkit (Och and Ney, 2003), which implements IBM models to generate the alignment of words or group of words between sentence pairs in both directions (Arabic-English and English-Arabic). We, then, symmetrized these two alignments using the grow-diag-final heuristic method. In the target language of the parallel corpus, we used IRSTLM (Federico *et al.*, 2008) to develop a 3-gram language model. In the development phase, we used an optimization algorithm based on the minimum error rate (MERT) (Och, 2003) to adjust the weights to each translation model in order to maximize translation quality.

4.3. Translation models

In this work, we have developed several Factored Translation Models and compared their performances. These factored models are obtained by integrating into the word representation of the Arabic language some linguistic information: lemma and/or PoS tag. This information provides a rich knowledge about the words in the corpus. The evaluation is carried out by comparing the different models with the standard model, which does not apply any pre-processing at the training corpus. The 4 translation models examined are:

- *Standard model:* this is the reference model in which we limit to exploiting the punctuation signs in the English/Arabic sentence to segment it, using the tokenizer script defined in the Moses system.
- *Factored models*: include 3 models:
  - The model (Word+Lemma) is the first model used. It is obtained by matching each word in the Arabic corpus with its diacritized lemma.
  - The second model (Word+PoS) is obtained by associating the PoS tag syntax label with each Arabic word.
  - The last model (Word+Lemma+PoS) is a combination of the last two and consists of exploiting both lemma and PoS tag information.

5. Results and discussion

In Table 2, we report the results for each of the four models tested on the development corpus, which is composed of 1000 sentences. The two measures used are the BLEU score (Papineni *et al.*, 2002) and the out-of-vocabulary word rate (OOV).

BLEU score		OOV (%)	
Standard model		28.51	28.6
Factored models	Word+Lemma	32.00	**19.6**
	Word+PoS	**32.22**	30
	Word+Lemma+PoS	32.09	20.3

Table 2. Evaluation results of the four translation models related to the development corpus

We notice that firstly, the three factored models have BLEU scores much higher than the standard model, and secondly, the BLEU scores of the three factored models are quite similar with a slight advantage for the model (Word+PoS).

Concerning the OOV indicator, the model (Word+Lemma) achieves the best performance with only 19.6% of words out of vocabulary. This says that the lemma information provides a better way to handle the scarcity or lack of some data in the corpus. However, the impact of reducing out-of-vocabulary words is not systematically reflected by the BLEU score since the models (Word+lemma+PoS) and (Word+PoS) have better BLEU scores than the model (Word+Lemma), while the latter has the better OOV indicator.

Regarding the assessment of these models on the test corpus that is composed of 1400 parallel sentences, we present, in Table 3, the BLEU score values.

BLEU score		
Standard model		21.48
Factored models	Word+Lemma	**31.08**
	Word+PoS	31.05
	Word+Lemma+PoS	**31.08**

Table 3. BLEU Score of the four translation models related to the test corpus

The results obtained are in accordance with the results obtained from the development corpus. Indeed, the BLEU scores of the three Factored Models are very close and much higher than that of the standard model.

We therefore conclude that the use of linguistic information has a major positive impact on the efficiency of translation systems.

6. Conclusion

In this work, we aimed to examine the impact of using linguistic knowledge on the performance of the Arabic/English translation system based on the PBSMT model. We experimented with some factored models obtained by injecting lemma and/or PoS tag information into the words of the source language (Arabic). The test results showed the benefits of using this information during the training phase in translation systems.

Following the results obtained, we plan to exploit other morphological information such as the roots or patterns of Arabic words. We also plan to use morphological information from the English words and test their impact on translation in both Arabic/English and English/Arabic directions. Finally, once we have expanded the training corpus, we will evaluate the potential impact of incorporating morphological information on the translation system based on neural model (NMT).

References

Fares A., Fan X., Alqwbani A., Al-Mansoub A. (2019).

Improved Arabic-Chinese machine translation with linguistic input features. Future Internet, Multidisciplinary Digital Publishing Institute, 11 (1): 22.

Ibrahim B., Zbib R., Glass J. (2008). Segmentation for English-to-Arabic statistical machine translation. Proceedings of ACL-08: HLT, pp. 153-56.

Koehn P., H. Hoang H. (2007). Factored translation models. Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), pp. 868-76.

Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N., Cowan B., Shen W., Moran C., Zens R., Dyer C., Bojar O., Constantin A., Herbst E. (2007). Moses: Open source toolkit for statistical machine translation. Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions, pp. 177–180. Association for Computational Linguistics.

Koehn P., Knowles R. (2017). Six Challenges for Neural Machine Translation. Proceedings of the First Workshop on Neural Machine Translation, pp. 28-39. Vancouver: Association for Computational Linguistics.

Koehn P., Och F. J., Marcu D. (2003). Statistical Phrase-based Translation. Proceedings of the 2003 Conference of the North American Chapter of the

Association for Computational Linguistics on Human Language Technology, Vol. 1, pp. 48-54. NAACL. Stroudsburg, PA, USA: Association for Computational Linguistics.

Manning C., Surdeanu M., Bauer J., Finkel J., Bethard S., McClosky D. (2014). The Stanford CoreNLP natural language processing toolkit. Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp. 55-60.

Marcello F., Bertoldi N., Cettolo M. (2008). IRSTLM: an open source toolkit for handling large scale language models. Ninth Annual Conference of the International Speech Communication Association.

Och F. J. (2003). Minimum Error Rate Training in Statistical Machine Translation. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, 1(1):160-67. Association for Computational Linguistics.

Och F. J., Ney H. (2003). A systematic comparison of various statistical alignment models. Computational linguistics, 29(1):19–51.

Papineni K., Roukos S., Ward T., Zhu W. J. (2002). BLEU: a method for automatic evaluation of machine translation. Association for Computational Linguistics, pp. 311-18.

Pasha A., Al-Badrashiny M., Diab M. T., El Kholy A., Eskander R., Habash N., Pooleery M., Rambow O., Roth R. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. LREC, pp. 1094-1101.

Rajeh H. A., Li Z., Ayedh A. M. (2016). A Novel Approach by Injecting CCG Supertags into an Arabic–English Factored Translation Machine. Arabian Journal for Science and Engineering, 41(8):3071-80.

Sennrich R., Haddow B. (2016). Linguistic Input Features Improve Neural Machine Translation. ArXiv.

Sreelekha S., Bhattacharyya P. (2017). Role of Morphology Injection in Statistical Machine Translation. arXiv preprint arXiv:1709.05487.

Ziemski M., Junczys-Dowmunt M., Pouliquen B. (2016). The United Nations Parallel Corpus v1. 0. Proceedings of the Tenth International Conference on Language Resources and Evaluation.

DiLiTAL'19, p. 143 - 160

Une nouvelle approche d'analyse sémantique des Tweets pour l'extraction des traits de personnalités

Marouane ECHHAIMI[1], Youness CHAABI[2], Tarik BOUJIHA[1]

[1]GEST, ENSA, Université Ibn Tofail, Kénitra, Maroc

[2]CEISIC, Institut Royal de la Culture Amazighe, Rabat, Maroc

{marouane.echhaimi, tarik.boujiha}@uit.ac.ma, chaabi@ircam.ma

Résumé.

L'utilisation des réseaux sociaux a énormément augmenté au cours de la dernière décennie, nous lisons et partageons des informations personnelles quotidiennement qui peuvent être analysées et utilisées de plusieurs manières, et ainsi, la reconnaissance des traits de personnalité des internautes sur les réseaux sociaux a attiré l'attention des chercheurs. La personnalité est un trait unique qui distingue les gens les uns des autres. Chacun a une façon différente de penser, de ressentir et de se comporter les uns envers les autres, et c'est ce qui affecte nos relations et interactions dans la société. L'une des théories les mieux documentées décrivant la variation des traits de personnalité serait le modèle des cinq facteurs, aussi connu sous le nom de « Big Five ». Les cinq grands traits sont l'Ouverture, la Conscience professionnelle, l'Extraversion, l'Agréabilité et le Névrosisme, soit OCEAN. La reconnaissance de la personnalité peut être utile dans divers domaines comme l'éducation, la sécurité, le marketing, le E-Learning, la médecine le recrutement de personnel etc. D'après des recherches antérieures, la personnalité peut être détectée par l'analyse automatique de texte.

Dans cet article, une approche multi-agents basée sur la mesure de similarité sémantique pour la reconnaissance des traits de personnalité à partir de l'analyse automatique des Tweets est proposée.

1. Introduction

Le réseautage social sur le Web a connu une croissance spectaculaire au cours de la dernière décennie. Les réseaux sociaux (Facebook, Twitter, etc)

sont devenus des médias populaires et largement utilisés pour la diffusion de l'information ainsi que pour les interactions sociales. Les activités des internautes sur les sites Web de réseautage social fournissent des renseignements précieux sur leurs personnalités. Ces personnalités fait référence aux différences individuelles dans la pensée, le comportement, les expériences, les sentiments, les opinions et les intérêts individuels (Kazdin, 2000). Connaître la personnalité est une façon de comprendre comment les diverses parties d'une personne se rejoignent dans leur ensemble, puisqu'il s'agit d'une combinaison de caractéristiques et de comportements d'un individu face à des situations différentes (Pratama, 2015). De plus, la personnalité peut influencer les choix des gens dans plusieurs domaines (Cantador, 2013). La personnalité affecte la façon dont nous interagissons avec les autres, nos relations et l'environnement qui nous entoure. La personnalité s'est avérée pertinente pour de nombreux types d'interactions et utile pour prédire la satisfaction au travail, le succès dans les relations professionnelles et même la préférence pour différents utilisateurs (Golbeck, 2011). La personnalité est importante pour plusieurs processus tels que : le recrutement du personnel, marketing en ligne, la conduite des thérapies psychologiques, le tutorat, l'enseignement aux étudiants et le conseil en santé, entre autres. Par conséquent, plusieurs applications pourraient bénéficier d'un aperçu de la personnalité. C'est pourquoi de nombreuses organisations s'intéressent à la connaissance de la personnalité des individus.

Dans la littérature de nombreuses approches d'identification des personnalités ont été proposées (Kafeza, 2014 ; Zheng, 2018 ; Gibson, 2018, Hawi, 2019). La problématique se présente quand ces approches ne prennent pas en considération la mesure de similarité sémantique, ce qui ne permet pas d'avoir des résultats concrets lors du calcul de la ressemblance de signification des textes. Les mesures de similarité de textes ont été utilisées dans de nombreux domaines. Par exemple, pour la classification de textes (Rocchio, 1971), la désambiguïsation du sens des mots (Lesk, 1986), la traduction automatique (Papineni *et al.*, 2002), l'identification de profils comportementaux sur le réseau social Facebook (Chaabi, 2019).

Dans cet article, un système multi-agents permet l'analyse des messages issus des réseaux sociaux et l'extraction des traits de personnalité des internautes en se basant sur la mesure de similarité sémantique. L'approche adopte la théorie des cinq facteurs « Big Five » (Paunonen, 2001). Il s'agit de nos jours du modèle de personnalité le plus reconnu par la communauté scientifique. Le modèle étant le fruit de plusieurs recherches indépendantes (Donnellan, 2006 ; Digman, 1997 ; De Raad, 2000).

2. Etat de l'art

Le volume de recherche sur la reconnaissance de la personnalité n'a cessé de croître au cours des dernières années (Celli *et al.*, 2014 ; Tkalcic *et al*., 2014). L'omniprésence des plateformes de médias sociaux a incité les chercheurs à s'orienter vers ces plateformes afin de trouver des informations utiles à utiliser pour la prédiction de la personnalité. De nombreuses études ont montré la corrélation entre la personnalité et le comportement en ligne (Hughes *et al.*, 2011 ; Bachrach *et al*., 2012). Quercia et al (Quercia *et al.*, 2011) ont été les premiers qui ont exploré la relation entre la personnalité et l'utilisation de Twitter ; ils ont également proposé un modèle pour déduire avec précision la personnalité des utilisateurs en se basant simplement sur trois compteurs publiquement disponibles sur les profils : les suivants, les suiveurs et les comptes listés (following, followers, and listed counts). De même, Jusupova et ses collaborateurs (2016) ont utilisé des données démographiques et des données sur l'activité sociale pour prédire la personnalité de l'enfant. Liu *et al.* (2016) ont proposé une approche fondée sur l'apprentissage en profondeur pour construire des systèmes hiérarchiques des représentations de mots et de phrases capables de déduire la personnalité des utilisateurs à partir de trois langues : Anglais, italien et espagnol. Van de Ven *et al.* (2017) ont fondé leurs analyses sur LinkedIn, un outil d'aide à la décision lié à l'emploi mais ils n'ont pas trouvé de fortes corrélations entre les traits de personnalité et l'utilisateur à l'exception de l'Extraversion. YouYou *et al.* (2014) ont démontré que les jugements informatisés au sujet d'un individu peuvent être plus précis que ceux faits par les amis, le conjoint, et même l'individu lui-même, si des données suffisantes sont disponibles.

3. Modèle du Big Five

Le modèle des cinq grands facteurs de la personnalité est un grand classique de la recherche en psychologie (Gosling, 2003 ; Barrick, 1991 ; Hough, 1992). Ces facteurs ne sont pas issus d'une perspective théorique mais ont été identifiés par des analyses du langage naturel et des tests psychologiques visant à identifier les traits de personnalité qui sont indépendants les uns des autres et décrivent la personnalité de la façon la plus complète possible. Les cinq grands traits sont l'Ouverture, la Conscience professionnelle, l'Extraversion, l'Agréabilité et le Névrosisme, soit OCEAN (Plaisant, 2010).

▪ **Ouverture** – Les personnes qui aiment apprendre de nouvelles choses et apprécient les nouvelles expériences marquent généralement un score

élevé en Ouverture. Cette dimension englobe des traits tels qu'être perspicace et imaginatif, et avoir de multiples centres d'intérêt.

▪ **Conscience professionnelle** – Les personnes qui ont un haut degré de conscience professionnelle sont fiables et ponctuels. Les traits incluent le fait d'être organisé, méthodique et rigoureux.

▪ **Extraversion** – Les extravertis puisent leur énergie dans l'interaction avec les autres, tandis que les introvertis la puisent au fond d'eux-mêmes. L'extraversion comprend des traits de type dynamique, loquace et assertif.

▪ **Agréabilité** – Ces individus sont amicaux, coopérants et doués de compassion. Les personnes ayant un score bas d'agréabilité peuvent être plus distantes. Parmi les traits, citons le fait d'être gentil, affectueux et sympathique.

▪ **Névrosisme** – On parle également de stabilité émotionnelle. Cette dimension porte sur la stabilité émotionnelle de la personne et son degré d'émotions négatives. Les personnes qui marquent un score élevé en Névrosisme sont souvent confrontées à une instabilité émotionnelle et à des émotions négatives. Les traits incluent le fait d'être d'humeur changeante et tendu.

4. L'approche proposée

L'approche proposée est d'analyser automatiquement le contenu des tweets. Suite à cette analyse, des personnalités pour chaque internaute est déterminé. Le défi consiste à identifier les traits de personnalité des utilisateurs de Twitter grâce à l'analyse automatique et sémantique du contenu des Tweets échangés. Chacun de ces Tweets subit une séquence de traitements (Chaabi *et al.*, 2019).

Dans cette section, nous présentons cinq grands profils de personnalité que nous avons identifiés et caractérisés selon différents critères. Afin de déterminer un profil, quatre traitements sont effectués. Après la récupération des Tweets, un second traitement réside dans la simplification des Tweets en supprimant les informations inutiles. Un troisième traitement consiste à une analyse linguistique qui effectue la normalisation des mots. Le quatrième permet de déterminer par une analyse sémantique des Tweets un profil de personnalités des Big Five pour chaque utilisateur dans le système. Voici l'architecture qui décrit le fonctionnement du système (Figure 1).

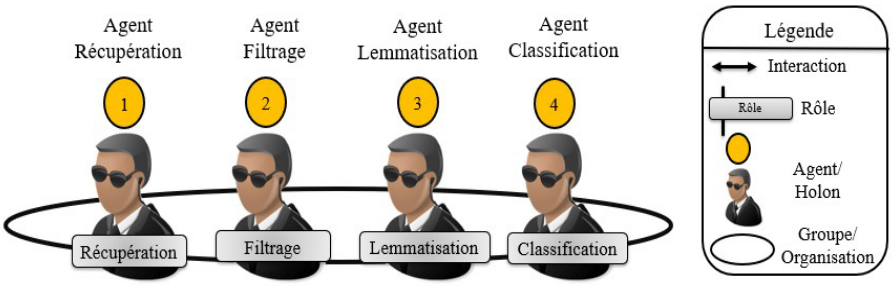


Figure 7. Architecture générale du système

4.1. Agent de récupération

Tout d'abord, un agent de récupération permet d'extraire les tweets (à l'aide de l'algorithme Tweepy (Roesslein, 2009) de Twitter, ainsi que de les préparer à différents traitements ultérieurs. Selon notre expérience, un premier traitement du corpus consiste en la correction des fautes d'orthographe et de grammaire. Des erreurs d'orthographe et de grammaire peuvent survenir dans l'analyse de texte, tant pour les humains que pour les logiciels. Un mot (ou une phrase mal orthographiée) peut changer complètement l'analyse. La correction orthographique et grammaticale est obtenue à l'aide d'un dictionnaire (corpus) associé à un algorithme qui prend en compte la variation du langage (conjugaison verbale, noms et adjectifs). Elle consiste à comparer les mots des Tweets avec le corpus, en tenant en compte le contexte des phrases. Néanmoins, l'utilité du correcteur orthographique et grammatical ne peut remplacer une vérification personnelle minutieuse.

4.2. Agent de filtrage

Une fois la tâche de récupération effectuée, un deuxième processus est chargé d'effectuer le traitement de filtrage. En effet, dans les messages textuels, de nombreux mots fournissent peu d'informations sur le message concerné. Ces mots sont automatiquement supprimés en utilisant des "mots vides" pour chaque langue (chaabi *et al.*, 2017).

Les mots qui apparaissent le plus souvent dans un corpus sont généralement des mots grammaticaux vides (Stop words) : articles, prépositions, mots de liaison, déterminants, adverbes, adjectifs indéfinis, conjonctions, pronoms et verbes auxiliaires, etc. Ces mots constituent une grande partie des mots d'un texte, mais sont malheureusement peu informatifs sur le sens d'un texte puisqu'ils sont présents sur l'ensemble des textes. Selon la loi de Zipf (Guiter, 1982), leur suppression lors du prétraitement des corpus permet de gagner du temps lors de la modélisation et de l'analyse.

4.3. Lemmatisation

Après l'étape de filtrage, le message fait l'objet d'une analyse linguistique qui effectue la normalisation des mots. Il permet la transformation des mots en leur forme canonique (Stemming) (Lovins, 1968). Ce traitement peut conduire à une réduction significative de la taille de l'échantillon du lexique (Chaabi *et al.,* 2015). Les règles de lemmatisation sont appliquées aux différents mots du corpus pour réduire les variantes morphologiques à une forme commune (mettre les verbes à l'infinitif, supprimer les formes plurielles...). Les variantes morphologiques d'un mot sont répertoriées dans le même lemme afin qu'ils puissent être analysés comme un seul élément (terme ou concept). En réduisant le nombre total de termes distincts, la lemmatisation diminue la complexité du texte analysé, et apporte donc d'importants bénéfices au système.

Dans de nombreuses langues, les mots apparaissent sous plusieurs formes. Par exemple, en Français, le verbe " marcher " peut apparaître comme " marcher ", " marche ", " marchait ", " marchent ", " marchaient ". La forme de base « marcher », que l'on pourrait chercher dans un dictionnaire, s'appelle le lemme du mot. La combinaison de la forme de base avec la partie de la parole est souvent appelée le lexème du mot.

4.4. Classification agent

L'agent de classification mesure la similitude sémantique d'un nouveau Tweet et détermine à quelle catégorie de personnalité (ouverture, conscience professionnelle, extraversion, agréabilité et névrosisme) il appartient à partir de la proportion des Tweets d'entraînement appartenant à cette catégorie.

4.4.1. Mesure de similarité sémantique

Dans de nombreux domaines de recherche tels que la psychologie, la linguistique, les sciences cognitives et l'intelligence artificielle, le calcul de la similarité sémantique entre les mots est une question importante (Wu, 1994). La similitude sémantique (ou proximité sémantique) est une métrique définie sur un ensemble de messages ou de termes, où l'idée de la distance entre eux est basée sur la ressemblance de leur significations ou contenus sémantiques (Zhu, 2017). D'autre part, par opposition à la similitude sémantique, on trouve le type de similitude qui peut être estimée en se basant sur les représentations syntaxiques des termes.

Des outils mathématiques sont utilisés pour estimer la force de la relation sémantique entre les unités de la langue, des concepts ou des instances, à travers une description numérique. Cette dernière est obtenue via la comparaison des informations à l'appui de leur signification ou la description de leur nature. La similarité sémantique peut être estimée par la définition d'une similitude topologique, en utilisant des ontologies pour définir la distance entre les termes/concepts (Hua, 2017). Par exemple, une métrique naïve pour la comparaison des concepts ordonnés dans un ensemble partiellement ordonné et représenté sous forme de nœuds d'un graphe orienté acyclique (par exemple, une taxonomie), serait le plus court chemin reliant les deux nœuds de concept.

La proximité sémantique entre les unités de la langue (par exemple des mots et des phrases) peut être également estimée en utilisant des moyens statistiques comme un espace vectoriel pour corréler des mots et des contextes textuels à partir d'un corpus de texte approprié.

4.4.2. Taxonomie

Le concept de similarité sémantique est plus spécifique que la parenté ou relation sémantique, car ce dernier inclut des concepts comme l'antonymie la méronymie, alors que la similarité ne les considère pas. Cependant, une grande partie de la littérature utilise ces termes de façon interchangeable avec des termes comme la distance sémantique (Jiang *et al*., 1997 ; Pedersen *et al*., 2007). Essentiellement, les notions de similarité sémantique, distance sémantique et proximité sémantique, apportent une réponse à la question suivante : "Quel est le degré de ressemblance entre le terme A et terme B?". La réponse à cette question est généralement un nombre compris entre -1 et 1, ou entre 0 et 1, où 1 signifie une similitude extrêmement élevée.

4.4.3. Mesure de la similarité topologique

Il existe essentiellement deux types d'approches qui calculent la similitude topologique entre les concepts ontologiques :

- Approche basée sur les arêtes : utilise les arêtes et leurs types comme source de données (Girardi *et al.*, 2016).
- Approche basée sur le contenu de l'information : les principales sources de données sont les nœuds et leurs propriétés (Wang *et al.*, 2017 ; Hao *et al.*, 2017).

4.4.4. Similitude sémantique

La similarité sémantique ou la relation sémantique est un concept de mesure de la proximité entre des termes ou documents dans le contexte de leur signification. Nous avons deux méthodes différentes pour calculer la similarité sémantique. L'une consiste à définir une similarité topologique, en utilisant l'ontologie pour définir une distance entre les mots. L'autre est basée sur l'utilisation des moyens statistiques tels que le modèle d'espace vectoriel pour établir une corrélation entre les mots et les contextes textuels à partir d'un corpus de texte approprié. Nous nous concentrons sur la première approche utilisant l'ontologie WordNet pour le calcul de similarité sémantique (Meštrović, 2016). Le calcul de similarité dans cette approche repose sur le fait que la similarité dépend des caractéristiques communes et distinctes des objets.

4.4.5. WordNet

WordNet est une ontologie lexicale pour la langue anglaise. Il s'agit d'un réseau sémantique développé par l'Université Princeton qui modélise le savoir lexical dans une hiérarchie taxonomique. WordNet contient trois bases de données : une pour les noms, une pour les verbes et l'autre pour les adverbes et les adjectifs. Les termes et concepts sont organisés en Synsets (Liste de termes ou concepts synonymes). La partie de base de WordNet est le Synset qui rassemble les synonymes d'un concept. Les Synsets sont liés dans certains modèles par des relations telles que : hypernymy (type de), meronymy (partie de) et antonymy (mot opposé) (Sudha, 2016). La similarité sémantique dans WordNet peut être calculée par deux méthodes : la longueur de chemin et le contenu d'information. La première méthode permet de calculer le nombre des nœuds ou de relations entre les nœuds dans la taxonomie. L'avantage de cette méthode est qu'elle n'est tributaire ni de la répartition statique du corpus, ni de la distribution des mots. Dans notre contexte, nous avons considéré uniquement deux concepts (relation et nom) dans la hiérarchie de WordNet. Nous utilisons WordNet 2.1 qui contient neuf hiérarchies de noms distinctes où parfois le chemin entre deux concepts peut ne pas exister (voir figure 2). Par conséquent, nous créons un nœud racine ("Entité" voir figure 2) qui reprend toutes les neuf hiérarchies données dans WordNet.

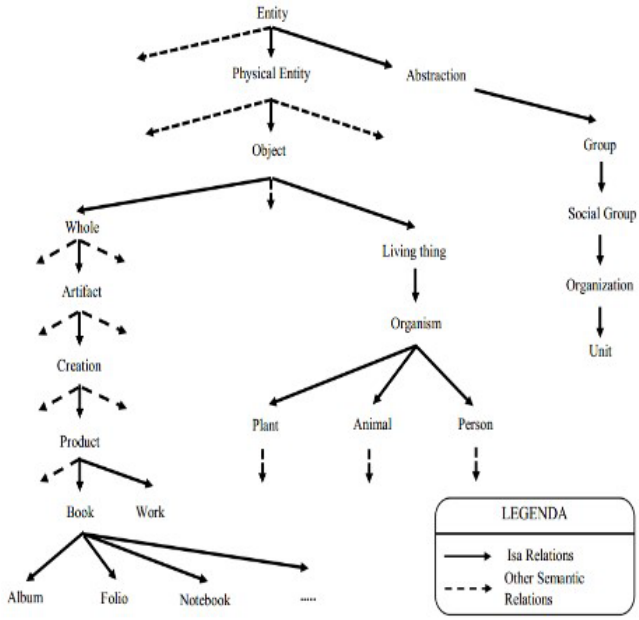


Figure 8. Extrait de la hiérarchie nominale de WordNet

4.4.6. Processus de mesure de similarité sémantique

L'agent classification permet de réaliser une séquence complète de traitement. Le processus de calcul de similarité sémantique est illustré à la figure 3. Ce processus se compose de trois phases :

- Phase 1 : Module de construction temporaire.
- Phase 2 : Module de calcul sémantique.
- Phase 3 : Procédures de mesure de similarité sémantique.

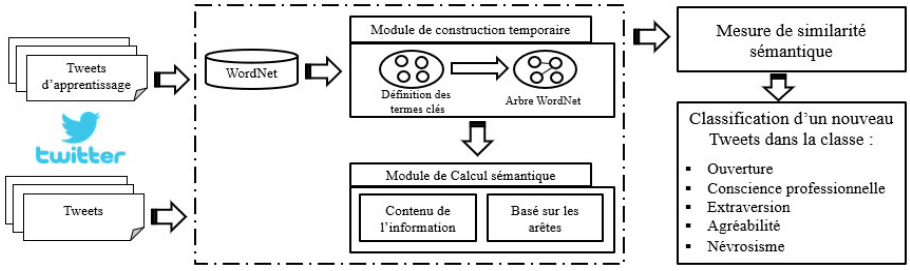


Figure 9. Diagramme de calcul de similitude sémantique

▪ **Phase 1 : Module de construction à terme**

L'objectif du module est de sélectionner tous les mots des Tweets qui existent sur WordNet et d'obtenir la relation entre ces mots. Nous utilisons WordNet pour générer une représentation de texte plus riche. Dans ce module, nous avons utilisé les hyperonymes fournis par WordNet comme des fonctionnalités utiles pour l'analyse de textes. L'objectif du module est de sélectionner tous les mots des tweets qui existent sur WordNet et d'obtenir la relation entre ces mots.

▪ **Phase 2 : Module de calcul sémantique**

D'après Philip Resnik (Resnik, 1995), les laboratoires Sun Microsystems présentent une alternative à la recherche de chemins via la notion de contenu informatif. C'est une mesure de spécificité attribuée à chaque concept dans une hiérarchie basée sur des éléments de preuve trouvés dans un corpus. Un concept avec un contenu informatif élevé est très spécifique, tandis que les concepts avec un contenu informatif faible sont associés à des concepts plus généraux. Le contenu d'information d'un concept est estimé en comptant la fréquence d'apparition de ce concept dans un grand corpus, ainsi que la fréquence de tous les concepts qui lui sont subordonnés dans la hiérarchie. La probabilité d'un concept est déterminée par une estimation du maximum de vraisemblance, et le contenu d'information est le log négatif de cette probabilité.

Resnik définit une mesure de similarité selon laquelle ces deux concepts sont liés sémantiquement proportionnellement à la quantité d'information qu'ils partagent. La quantité d'informations partagées est déterminée par le contenu informatif du concept le plus bas dans la hiérarchie qui recouvre les deux concepts donnés.

La similarité des mots basée sur le contenu informatif :

- dépend encore de la structure du thésaurus ;
- améliore l'approche basée sur le chemin en utilisant des normalisations basées sur la profondeur de la hiérarchie ;
- représente la distance associée à chaque arête ;
- ajoute des informations probabilistes dérivées d'un corpus ;

La probabilité que le mot aléatoire soit une instance du concept est (1) :

$$p(c) = (\sum w \in w_{(c)} \text{count(w)}) / N \quad (1)$$

où : words(c) est l'ensemble des mots subsumés par le concept c

N est le nombre de mots dans le corpus et dans le thésaurus
P(root)=1 puisque tous les mots sont subsumés par le concept racine

Plus le concept est bas dans la hiérarchie, plus la probabilité est faible.

Nous avons besoin de deux autres définitions :

1) Le contenu informatif d'un concept (2) :

$$IC(c) = -\log P(c) \quad (2)$$

Théorie de l'information de base

2) Le subsumer commun le plus bas : LCS (c1, c2)

Il s'agit du nœud le plus bas dans la hiérarchie qui est un hyperonyme de c1 et c2

La mesure Similarité de Resnik est (3):

$$\text{simResnik}(c1, c2) = -\log \text{P}(\text{LCS}(c1, c2)) \quad (3)$$

Elle estime la quantité commune d'information entre les mots en utilisant le contenu informatif du subsumer commun le plus bas.

▪ **Phase 3 : Mesures de similarité sémantique**

Les vecteurs sémantiques pour T_1 et T_2 peuvent être formés à partir de statistiques T et de corpus. Le processus de dérivation des vecteurs sémantiques pour T_1 (4) :

Mot w, définir

$$Sim(W_1, W_2) = max_{c1,c2}[smin(c1, c2)]$$

$$Sim(T1, T2) = \sum_{i=1}^{n} \left(\frac{sim(Wi, Wi+1}{n}\right) \quad (4)$$

Nous obtenons des valeurs de mesure de similarité sémantique pour Resnik entre le Tweet $_1$ et le Tweet $_2$ (5) :

$$\text{- Sim Resnik } (T_1, T_2) = \text{valeur}_2 \quad (5)$$

Les Tweets sont composés de mots, il est donc raisonnable de représenter un Tweet en utilisant les mots qu'il contient.

Contrairement aux méthodes traditionnelles qui utilisent une liste de mots précompilée contenant des centaines de milliers de mots, notre méthode ne forme dynamiquement des vecteurs sémantiques que sur la base des Tweets comparés. Les recherches récentes en analyse sémantique sont généralement adaptées pour extraire automatiquement un vecteur sémantique de mots pour une phrase (Resnik, 1995). Avec deux Tweets T_1 et T_2, un ensemble de mots est formé avec (6) :

$$T = T1 \cup T2 = \{W_1, W_2, \ldots\ldots, W_n\} \quad (6)$$

L'ensemble de mots T contient tous les mots distincts de T_1 et T_2. La morphologie flexionnelle peut faire apparaître un mot dans un message sous différentes formes qui ont une signification particulière pour un contexte spécifique. Pour cette raison, nous utilisons le mot forme tel qu'il apparaît dans le Tweet.

5. Expérimentations et résultats

Le modèle objet de ce travail a pour finalité d'automatiser l'identification des personnalités des internautes par une analyse sémantique des Tweets. Dans ce sens, nous avons procédé à une étude comparative entre l'évaluation humaine et celle produite : résultat de notre modèle.

Nous avons mené des expériences d'analyse intuitive sur des Tweets d'internautes à partir de note corpus de test. Nous nous intéressons ici à l'analyse qualitative et quantitative réalisées auprès de 3 experts. Un corpus des Tweets fut élaboré à partir d'un échantillon de 10 internautes. Notre corpus est constitué de 100 Tweets publiés par chaque utilisateur. L'analyse intuitive des Tweets consiste à : (1) associer une personnalité à chaque internaute, (2) identifier les actes de langages qui déterminent la personnalité.

Pour identifier les traits de personnalités, nous avons choisi la base de données MyPersonnality comme base d'apprentissage pour notre système.

5.1. Base de données Mypersonnality

Afin de tester notre approche, nous nous appuyons sur l'ensemble de données créé par le projet MyPersonality. L'ensemble de données a été recueilli à des fins de recherche par David Stillwell et Michael Kosinski à l'aide d'une application Facebook qui a administré un test de personnalité et a recueilli un large éventail d'informations personnelles et d'activités à partir du profil des utilisateurs de Facebook, sous leur consentement. L'application MyPersonality a été active de 2007 à 2012 et a collecté une énorme quantité de données.

Nous fondons notre étude sur un échantillon de l'ensemble de données original de MyPersonality qui a été rendu public (Celli *et al.*, 2014). Le modèle contient 9913 mises à jour de statut en anglais de 250 utilisateurs (anonymisées), annotées de scores de traits de personnalité et de statistiques de base décrivant le réseau social de l'utilisateur.

5.2 Corpus de test

Pour avoir une base de test appropriée, on a pu récupérer à l'aide de l'algorithme Tweepy (Roesslein, 2009) de Twitter plus que 1.5 millions de Tweets ; on a ensuite choisi 10 utilisateurs ayant au moins 100 Tweets par personne, la base est en format CSV et chaque ligne comporte 6 champs :

0 - la polarité du Tweet

1 - l'id du Tweet

2 - la date du Tweet

3 - la requête. S'il n'y a pas de requête, alors cette valeur est NO_QUERY

4 - l'utilisateur qui a tweeté

5 - le texte du Tweet

	Agréabilité	**Conscience professionnelle**	**Extraversion**	**Névrosisme**	**Ouverture**
Bigeny	34.33 %	14.91 %	13 %	17.93 %	19,83 %
Broooook	46.41 %	4.57 %	10,11 %	22.22 %	16.69 %
Djalf	31.62 %	0.68 %	12, 42 %	33.51 %	21.77 %
Dogboo	06.12 %	24.56 %	17.26	16.22 %	35.84 %
Dontforgetchao	33.08 %	09.00 %	46.62 %	20.00 %	6.57 %
Dutchrudde	60.15 %	12.14 %	05.98 %	17.73 %	04.00 %
Hollyalyxfinc	07.56 %	51.12 %	07.00 %	28.00 %	06.32 %
HollywoodTre	11.00 %	08.00 %	56.27 %	14.00 %	10.73 %
Jayme198	08.00 %	13.16 %	71.06 %	02.78 %	05.00 %
JBnVFCLover78	07.13 %	04.22 %	08.65 %	80.00 %	02.56 %

Tableau 2. Résultat de l'exécution de notre système sur le corpus de test

Nous avons comparé le texte des Tweets avec notre base d'apprentissage MyPersonality en tenant en compte la mesure de similarité sémantique afin d'extraire les traits de personnalité plus précisément. Le tableau ci-dessus représente le résultat de l'exécution de notre système sur le corpus de test :

Si par exemple, on prend le dernier utilisateur « JBnVFCLover78 » sur le tableau (1), on remarque évidemment que son score en névrosisme est très élevé, ce qui montre que cette personne n'est pas stable émotionnellement, voici quelques extraits des Tweets de cet utilisateur :

- "I'm pretty **pissed** at pretty much everything right now"
- "I was **sad** yesterday"
- "Now I'm **depressed** again..ugh!!! HELP !!! "
- "Im so **sad** right now"
- "Twitter is making me **sad**"
- "I need some cheering up guys"
- "yeah it's not going to be ok. I just can't stop thinking of him"
- "i still **shed tears** all the time and I miss u"
- "i need some awesome songs to cheer me up"
- "the **worst** thing possible happened"
- "im **mad** at myself now"
- "I'm **confused** too"
- "I just feel like banging my head against something"
- "that's **annoying**"
- "that makes me **sad**"
- "that gets me **mad** too"

On constate clairement l'utilisation des expressions exprimant la mélancolie et l'instabilité émotionnelle de cet utilisateur ce qui explique le taux élevé en névrosisme.

Après comparaison des résultats, nous avons trouvé une parfaite concordance entre les résultats intuitifs des experts et les résultats de notre système.

6. Conclusion et perspectives

Les traits de personnalité ont un impact sur tous les aspects de l'être humain. C'est un acteur important dans la prise des décisions et la réussite. Connaître la personnalité des gens est utile pour comprendre divers aspects tels que les comportements collectifs et les préférences des gens, ainsi la personnalité pourrait être utile pour sélectionner des candidats pour un emploi ou pour planifier un marketing ciblé ou pour des raisons de sécurité. Nous nous intéressons à la détection de la personnalité en tenant en compte la mesure de similarité sémantique des Tweets avec notre base d'apprentissage basée sur le modèle Big Five. L'exécution de notre approche sur un échantillon de données de la base de test récupérée via l'API Tweepy de Twitter a permis d'avoir des résultats très prometteurs. Cependant, nous espérons d'améliorer notre système, premièrement, diminuer le temps d'exécution, même si on a limité notre test sur 10 utilisateurs de Twitter, le système a pris plusieurs heures pour s'exécuter. Deuxièmement, une base de test avec plus d'informations sur les utilisateurs permettra d'avoir de meilleurs résultats.

Références

Kazdin E. (2000). Encyclopedia of Psychology: 8 Volume Set. American Psychological Association and Oxford University Press. 2000.

Pratama B. Y., Sarno R. (2015). Personality classification based on Twitter text using Naïve Bayes, KNN and SVM. International Conference on Data Softw. Eng., pp. 170-174.

Cantador I., Fernández-Tobías I., Bellogín A. (2013). Relating personality types with user preferences in multiple entertainment domains, CEUR Workshop Proc., Vol. 997.

Golbeck I., Robles C., Edmondson M., Turner K. (2011). Predicting personality from twitter, IEEE Int. Conf. Privacy, Secur. Risk Trust IEEE Int. Conf. Soc. Comput. PASSAT/SocialCom 2011, pp. 149-156.

Kafeza E., Kanavos A., Makris C., Vikatos P. (2014). T-PICE: Twitter personality based influential communities' extraction system. IEEE International Congress on Big Data, pp. 212-219.

Zheng Y. (2018). Identifying Dominators and Followers in Group Decision Making based on The Personality Traits. In IUI Workshops.

Hawi N., Samaha M. (2019). Identifying commonalities and differences in personality characteristics of Internet and social media addiction profiles: traits, self-esteem, and self-construal. Behaviour & Information Technology, 38(2):110-119.

Paunonen S. V., Ashton, M. C. (2001). Big five factors and facets and the prediction of behavior. Journal of personality and social psychology. Vol. 81, n° 3.

Donnellan M. B., Oswald F. L., Baird B. M., Lucas R. E. (2006). The miniIPIP scales: tiny-yet-effective measures of the Big Five factors of personality. Psychological assessment. Vol. 18, n° 2.

Digman J. M. (1997). Higher-order factors of the Big Five. Journal of personality and social psychology, 73(6):12-46.

De Raad B. (2000). The Big Five Personality Factors: The psycholexical approach to personality. Hogrefe & Huber Publishers.

Celli F., Lepri B., Biel J., Gatica-Perez D., Riccardi G. (2014). The workshop on computational personality recognition 2014. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA. pp. 1245-1246.

Tkalcic M., de Carolis B., de Gemmis, M., Odi´c A., Košir A. (2014). Preface: EMPIRE 2014. 2nd Workshop Emotions and Personality in Personalized Services. Proceedings of the 22st Conference on User Modeling, Adaptation, and Personalization, Aalborg, Denmark.

Hughes D., Rowe M., Batey M., Lee A. (2011). A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. Comput. Hum. Behav, Vol. 28, pp. 561-569.

Bachrach Y., Kosinski, M., Graepel, T., Kohli, P., Stillwell, D. (2012). Personality and patterns of Facebook usage. In Proceedings of the 4th Annual ACM Web Science Conference, Evanston, IL, USA.

Quercia D., Kosinski M., Stillwell D., Crowcroft J. (2011). Our twitter profiles, our selves: Predicting personality with twitter. IEEE Third International Conference on Privacy, Security, Risk and Trust and International Conference on Social Computing (SocialCom), Boston, MA, USA, 9-11 October 2011, pp. 180-185.

Jusupova A., Batista F., Ribeiro R. (2016). Characterizing the Personality of Twitter Users based on their Timeline Information. In Proceedings of the Atas da 16 Conferência da Associacao Portuguesa de Sistemas de Informação, Porto, Portugal, 22-24 December 2016, Vol. 16, pp. 292-299.

Liu F., Perez J., Nowson S. (2016). A Language-independent and Compositional Model for Personality Trait Recognition from Short Texts. arXiv 2016, arXiv:1610.04345.

Van de Ven N., Bogaert, A., Serlie, A., Brandt, M.J., Denissen, J.J. (2017). Personality perception based on LinkedIn profiles. Vol. 32, pp. 419-429.

YouYou W., Kosinski, M., Stillwell, D. (2014). Computer-based personality judgments are more accurate than those made by humans. Proc. Natl. Acad. Sci. USA. Vol. 112, pp. 1036-1040.

Gosling, S. D., Rentfrow, P. J., & Swann Jr, W. B. (2003). A very brief measure of the Big-Five personality domains. Journal of Research in personality, 37(6):504-528.

Barrick M. R., Mount M. K. (1991). The big five personality dimensions and job performance: a meta-analysis. Personnel psychology. 44(1):1-26.

Hough L. M. (1992). The'Big Five'personality variables--construct confusion: Description versus prediction. Human performance, 5(1-2):139-155.

Plaisant O., Guertault J., Courtois R., Réveillère C., Mendelsohn G. A., John, O. P. (2010). Histoire des « Big Five » : OCEAN des cinq grands facteurs de la personnalité. Introduction du Big Five Inventory français ou BFIFr. In Annales Médico-psychologiques, revue psychiatrique. 168(7):481-486. Elsevier Masson.

Roesslein J. (2009). Tweepy Documentation. Online http://tweepy. Readthedocs. io/en/v3, 5.

Chaabi Y., Lekkidioui K., Jebbor F., Messoussi R. (2017). Determination of Distant Learner's Sociological Profile Based on Fuzzy Logic and Naïve Bayes Techniques. International Journal of Emerging Technologies in Learning (iJET), 12(10):56-75.

Guiter H., Arapov M. V. (Eds.). (1982). Studies on Zipf's law, Vol. 16. Brockmeyer.

Lovins J. B. (1968). Development of a stemming algorithm. Mech. Translat. & Comp. Linguistics, 11(1-2):22-31.

Chaabi Y., Messoussi R., Hilaire V., Ruichek Y., Lekdioui K., Touahni R. (2015). Design of an Intelligent System to Support Tutors in Learning Communities Using Multi-Agent Systems and Fuzzy Logic. International Review on Computers and Software (IRECOS), 10(8):845-855.

Wu Z., Palmer M. (1994). Verbs semantics and lexical selection. The 32nd annual meeting on Association for Computational Linguistics, pp. 133-138. Association for Computational Linguistics.

Zhu G., Iglesias, C. A. (2017). Computing Semantic Similarity of Concepts in Knowledge Graphs. IEEE Transactions on Knowledge and Data Engineering, 29(1):72-85.

Hua W., Wang Z., Wang H., Zheng K., Zhou X. (2017). Understand Short Texts by Harvesting and Analyzing Semantic Knowledge. IEEE transactions on Knowledge and data Engineering, 29(3):499-512.

Jiang J. J., Conrath D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008.

Pedersen T., Pakhomov S. V., Patwardhan S., Chute, C. G. (2007). Measures of semantic similarity and relatedness in the biomedical domain. Journal of biomedical informatics, 40(3):288-299.

Girardi D., Wartner S., Halmerbauer G., Ehrenmüller M., Kosorus H., Dreiseitl, S. (2016). Using concept hierarchies to improve calculation of patient similarity. Journal of Biomedical Informatics, Vol. 63, pp. 66-73.

Wang H. Y., Ma W. Y. (2017). Integrating Semantic Knowledge into Lexical Embeddings Based on Information Content Measurement. EACL 2017.

Hao W. U., HUANG, H. (2017). Efficient Algorithm for Sentence Information Content Computing in Semantic Hierarchical Network.

Meštrović A., Calì, A. (2016). An ontology-based approach to information retrieval. In Semanitic Keyword-based Search on Structured Data Sources, pp. 150-156. Springer, Cham.

Sudha S. (2016). WordNet–An On-line Lexical Database.

Resnik P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada, pp. 448-453.

Chaabi Y., Lekkidioui K., Boumediane M. (2019). Semantic Analysis of Conversations and Fuzzy Logic for the Identification of Behavioral Profiles on Facebook Social Network. International Journal of Emerging Technologies in Learning. Vol. 14, n° 7.

Rocchio J. (1971). Relevance feedback in information retrieval. The Smart retrieval system-experiments in automatic document processing, pp. 313-323.

Lesk M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In Proceedings of the 5th annual international conference on Systems documentation, pp. 24-26.

Papineni K., Rouko, S., Ward T., Zhu W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics, pp. 311-318. Association for Computational Linguistics.

Salton G., Singhal A., Mitra M., Buckley C. (1997). Automatic text structuring and summarization. Information processing & management, 33(2):193-207.

ⵙⵓⵛⵖ++ⵓ ⵖⴰⵎⵙⵓ ⵕⴰ +ⵙⵍⵖⵓⵙⵍⵙⵉ ⵕ ⵙⵣⵓⵕⵙ ⵔ ⵖⵣⵓⵕ+ⵙ +ⵕⵍⵕⵓⵎⵕ +ⵙⵓ ⵓⵍⵕ+ ⵏ +ⵙⵍⵕⵓⵍ+ ⵅⵣ ⵖⵛⵟⵕⵍⵕⵓⵍ ⵕⵓⵏⵙⵎⵓ ⴰ ⵖⵓⵛⴽⵎ ⵏ +ⵖ+ⵎⵕⵟⵙⵉ ⵎⵎⵙ ⵙⵓⵏⵛⵕⵎⵕ ⵍⵕⵛⵛⵕⵓ ⵏ +ⵟⵖⵓⵙⵍⵙⵉ ⵏ +ⵓⵏⵛⵕⵎⵕⵟ+ ⴰ ⵙⵛⵕⵅⵓⵕⵍⵏ ⵏ ⵖⵏⵟⵛⵙⵓ ⴰ ⵖⵛⵟⵕⵍⵕⵓⵍⵕ ⵕⵣⴴ ⵖⵓⵕⵕ ⵅ +ⵟⵖⵓⵙⵍⵙⵉ ⵏ +ⵓⵏⵛⵕⵎⵕⵟ+, ⵖⵓⵓⵕⵏ ⵏ 16 ⴰ 17 ⴰⵖⵏⵕⵏⵓⵙⵓ 2019 ⵅ +ⵟⵖⵙⵍⵕⵏ+ ⵏ +ⵛⵕⵓⵓⵕⵏⵙⵉ ⵏ ⵍⵏⴰⵕ.

+ⵙⵍⵖⵓⵙⵍⵙⵉ ⵕⴰ ⵓⵛⵖⵏ+ ⵕⴰⵅⵕⵍ ⵏ ⵙⵛⵖⵣⵕⵟⵏ ⴰ ⵙⵛⵓⵣⵖⵟⵏ ⵙⵏⵕⵛⵖⵓⵏ ⴰ ⵙⵅⵓⵟⵎⵕⵏ ⵅ ⵟⵙⵅⵓ ⵏ +ⵛⵕⵓⵓⵕⵏⵙⵉ +ⵙⵛⵖⵄⵄⵖⵏⵙⵉ +ⵖⵓⵏⵙⵓⵙⵉ ⵣⵅ +ⵖ+ⵎⵕⵟⵙⵉ +ⵙⵅⵓⵛⵕⵏⵙⵉ.

* * *

قدمت الأعمال المنشورة في هذا الكتاب في إطار الدورة الثانية للندوة حول التنوع اللغوي ومعالجة اللغات الطبيعية التي تم تنظيمها من طرف مركز الدراسات المعلوماتية وأنظمة الاعلام والاتصال، بشراكة مع مختبر البحث في الدراسات المعلوماتية، يومي 16 و17 دجنبر 2019 بكلية العلوم بوجدة.

يسلط هذا الكتاب الضوء على مساهمات الخبراء والباحثين الوطنيين والدوليين العاملين في مجال العلوم الرقمية المطبقة على اللغات الطبيعية.

* * *

Le présent ouvrage rassemble les travaux exposés au cours de la 2e édition de la conférence sur la diversité linguistique et TAL (DiLiTAL), organisée les 16 et 17 décembre 2019 à la Faculté des Sciences d'Oujda, par le Centre des Etudes Informatiques, des Systèmes d'Information et de Communication et le Laboratoire de Recherche en Informatique.

Cet ouvrage met en exergue les contributions des experts, des chercheurs nationaux et internationaux qui œuvrent dans le domaine des sciences du numérique appliquées aux langues naturelles.

50 DH